Validating Next Generation Sequencing for Meiofaunal Community Analysis and Interaction Prediction

Ben Nichols

Submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

School of Engineering

College of Science and Engineering
University of Glasgow

October 2015

© Ben Nichols

Abstract

Advances in DNA sequencing technologies, particularly the advent of next generation sequencing (NGS) platforms, have revolutionised the field of metagenomics and allowed great progress to be made in the way that microbial communities are analysed. However, the wealth of data now available thanks to these advancements has made the possibilities far more numerous than just the obvious applications, with a wide variety of novel and diverse studies conceivable. The technologies themselves have also created further areas for research as better methods of handling the, often overwhelming in quantity and misleading in content, data are sought.

The analysis carried out in this thesis demonstrates the wide range of study possible stemming from two experiments involving the sequencing of meiofauna DNA. The first of these involves community analysis of marine benthic meiofauna with particular emphasis on diversity and distribution. The second experiment involves the sequencing of pooled nematode samples in order to investigate the effects of sample richness and species relatedness on the generation of chimeric reads in sequencing data.

It is shown that the data generated from these two experiments can be used to help formulate an algorithm to simulate PCR and therefore assist the generation of realistic noisy NGS data. These data can, in turn, be used to generate a simulated *in silico* microbial community for analysis, the results of which reveal insights into the accuracy of chimera detection software and the reliability of metagenetic community analyses. Worryingly, these results suggest that findings from similar *in vitro* studies are not as reliable as originally perceived.

The same experimental data may also be used to investigate interactions between meiofauna species based on the incidental presence of prey species highlighted from the sequencing of individual meiofauna organisms. It is shown that these data can be used to accurately predict a nematode's feeding type without having to examine the organism directly. It is also shown that there is no correlation between this method of inferring interactions between species and other methods which have been used in the past. This suggests that the earlier methods are inadequate when used for the detection of feeding interactions.

Acknowledgements

I would like to thank my two supervisors, Chris Quince and Bill Sloan, firstly for giving me the opportunity to undertake my PhD and also for their support and encouragement throughout my time at Glasgow University. I would also like to thank the Unilever Research and Development department for supporting my project.

Thanks to Vera Fonseca and Simon Creer from Bangor University for their invaluable collaboration and experimental work, the data from which helped to make many of the analyses in this thesis possible. Thanks also to Tom Moens from Ghent University for his expert help in assigning feeding types to marine nematodes for the construction of interaction networks.

Finally, thank you to the postgraduate and postdoctoral researchers with whom I worked closely at the University of Glasgow. Umer, Keith, Jillian, Zofia, Melanie, Sarah, Siding, Dustin, Rick and Stephanie, thanks for the help and for creating a pleasant environment in which to work.

Declaration

I declare that, except where explicit reference is made to the contribution of others, this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Ben Nichols, Glasgow, April 2015.

Dedicated to my parents for always supporting me.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Introduction

Most of the work presented in this thesis involves computational analysis of next generation DNA sequencing (NGS) data after the actual experimental work has already taken place. However, knowledge of the technology and methods used to generate these data is critical for the understanding of this work. For example, strategies for the removal of noisy sequences will, naturally, make use of information about how sequencing noise is generated. This chapter covers all relevant background knowledge for the complete understanding of the analyses presented throughout this thesis.

The fundamental aspects of DNA sequencing are summarised in Section 1.2. Section 1.3 describes computational sequence representation and alignment, which are used in many of the bioinformatic tools that are featured in later chapters. Issues related to sequencing noise are covered in Section 1.4. Aspects of microbial community analysis, a subject which benefits greatly from the use of NGS data, are outlined in Section 1.5. Section 1.6 then goes on to explain how these ideas are brought together to produce the studies described in this thesis.

1.2 DNA and Sequencing

1.2.1 DNA

The underlying feature of all of the analysis described in this thesis is the study of Deoxyribonucleic acid (DNA) sequences. DNA is an essential component of all living organisms and DNA molecules are generally stored in the nuclei of eukaryotic organisms and the cell

cytoplasm of archaea and bacteria. The function of DNA is to encode the genetic information of an organism, which it is able to do because of its structure.

A DNA molecule is, fundamentally, comprised of two complementary strands made up of four different nucleotides, or nucleobases (bases) - Adenine (A), Cytosine (C), Guanine (G) and Thymine (T). These bases are arranged sequentially and it is the order of this sequence which maps the biological information held within the DNA molecule. The two complementary strands both hold the same information expressed differently. As shown in Table 1.1 each of the four nucleotides has a complementary partner which takes the same position on the complementary strand. Thus if the leading DNA strand has the sequence *ACGT* then the complementary strand will have the sequence *TGCA*.

Nucleotide Name	IUPAC Code	Type	Complementary Nucleotide
Adenine	A	Purine	T
Cytosine	C	Pyrimidine	G
Guanine	G	Purine	C
Thymine	T	Pyrimidine	A

Table 1.1: The four nucleotides that comprise a DNA molecule.

DNA molecules are divided into regions called *genes* which, typically, control specific functions of the organism to which the DNA belongs. Genes from a particular organism can be compared to the equivalent genes of other organisms to detect biological differences which may, or may not, have an affect on the functionality or appearance of the organisms in question.

The deceptively simple structure of a DNA molecule is what makes it ideal for storing information and it is also this simplicity that allows this information to be readily analysed after it has been read because it can be stored as long strings of the four letters, A, C, G and T (the sequence of the complementary strand need not be recorded because it can be inferred directly from the main sequence). A common way to store the information is in *fasta* format in a text file which simply lists each sequence in a dataset, a format which lends itself well to computational analysis. From these files, to give two examples, complex analysis of microbial communities based on their DNA sequences and simulations of biological processes can be performed away from the laboratory.

The actual “reading” of the genetic information (by extracting and sequencing the DNA) is more complex.

1.2.2 DNA Sequencing

DNA sequencing is the process of documenting the sequences of bases from DNA molecules. Historically, there have been many ways of doing this but the studies in the later chapters deal mainly with next generation sequencing techniques. One of the earlier methods that is also referred to is Sanger sequencing for which there is an overview in Section 1.2.2.

Primers

A DNA Primer is a short strand of DNA which is used to instigate DNA synthesis. It is required because a new DNA strand can only be formed by attaching bases sequentially to the end of an existing strand of DNA. Primers are used during DNA sequencing to replicate the existing DNA that is undergoing analysis.

PCR Primers

In PCR (see Section 1.2.3), primers are generally used in pairs consisting of a *forward primer* and a *reverse primer*. The primers highlight a targeted region of DNA by each attaching to a conserved section either side of it - the primers attach because they are chosen to be a complementary match (or, in practice, close to a match) to these conserved sections. The two ends of a DNA strand are labelled $3'$ and $5'$. The forward primer will attach to the leading strand of DNA which always synthesises in the $5' \rightarrow 3'$ direction and the reverse primer will attach to the complementary strand which synthesises in the opposite direction. Thus, the desired region of DNA is highlighted as shown in Figure 1.1.

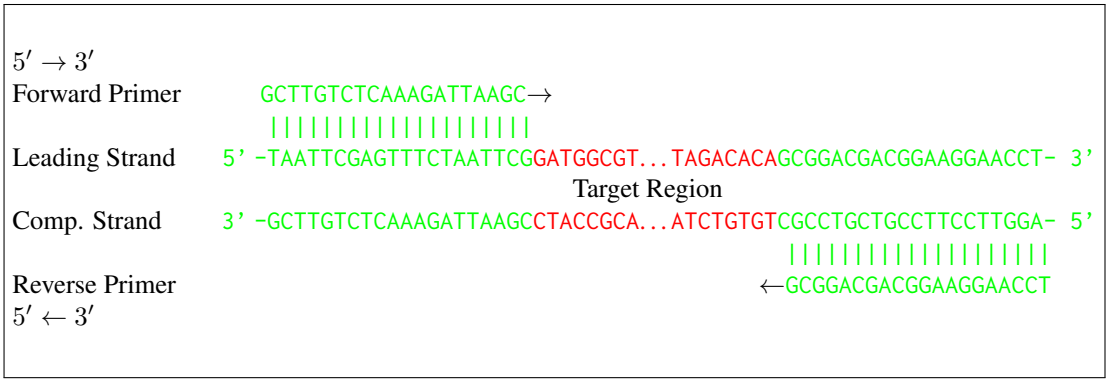

Figure 1.1: Forward and reverse PCR primers. The forward primer binds with the leading strand and the reverse primer binds with the complementary strand to form two duplicated molecules. Notation referring to the direction of each primer is relative to the leading strand.

Sanger Sequencing

Sanger sequencing (1) requires the use of a *DNA polymerase* which is an enzyme used to create DNA molecules by assembling nucleotides. DNA polymerase is often used to replicate an existing DNA molecule and is the essential ingredient in PCR which is described in Section 1.2.3. In addition, Sanger sequencing uses normal deoxynucleoside triphosphates (dNTPs: dATP, dCTP, dGTP and dTTP) which contain the bases A, C, G and T and can be used for DNA synthesis in conjunction with modified, chain-terminating di-deoxynucleotide triphosphates (ddNTPs: ddATP, ddCTP, ddGTP and ddTTP) which cause synthesis to cease when they are incorporated at the end of a chain. This sequencing method also requires the use of a primer and a single-stranded DNA template.

Four separate experiments are conducted, all of which include dNTPs containing all four bases but only the ddNTP containing one specific base. That is, there will be one experiment which only uses ddATP, one that uses ddCTP, one that uses ddGTP and one that uses ddTTP. DNA extension is initiated and will terminate, in each experiment, when a ddNTP molecule is incorporated into the chain. After a large number of DNA extensions have occurred in, for example, the experiment using ddATP it will be apparent at which positions on the DNA strand the base T (the complementary base to A) appears. The data from all four experiments can then be combined to give a complete DNA sequence.

The ddNTPs are radioactively or fluorescently labelled and the resultant sequence can be visualised using auto-radiography or UV light.

1.2.3 Next Generation DNA Sequencing

Low cost and high throughput Next Generation Sequencing methods have allowed analysis to take place that was, previous to the availability of NGS platforms, either unrealistic or completely impossible. NGS methods parallelise the process which allows thousands or millions of sequences to be recorded concurrently.

The data used for analysis in this thesis were generated using pyrosequencing and for this reason, the procedure for this method has been outlined in this section. Sequencing by synthesis (Illumina sequencing) has also been covered to allow comparison between pyrosequencing and another NGS method.

Although NGS methods are highly advantageous in terms of time and cost, there is a trade-off in the level of accuracy of the data produced. Some of these concerns are addressed in Section 1.4.

PCR Amplification

In all NGS methods, the DNA to be analysed must first be amplified. To prepare the sample for sequencing, an amplification step is carried out using Polymerase Chain Reaction (PCR). Thermal cycling is used to repeatedly melt and cool the DNA. When a strand of DNA is copied, this copy can then also be copied; this leads to an exponential amplification effect. PCR is used to amplify a particular target region of the DNA - this is selected using primers (small pieces of DNA, complementary to the target region).

The process typically involves 20-40 cycles of the following steps (2^{40} gives approx 10^{12} copies):

1. Denaturation – this step takes place at temperatures between 94 and 98°C for around 20 to 30 seconds. Hydrogen bonds are broken to split the DNA into two strands.

2. Annealing – the temperature is reduced to 50-65°C. The primers bind to both single strands of DNA. Hydrogen bonds are only able to form when there is a close match, ensuring that the primers are annealed to the correct region.

3. Extension – the temperature is adjusted depending on the polymerase used. Nucleotides are attached to complete the DNA strands. These strands can now be copied in the same way as the original.

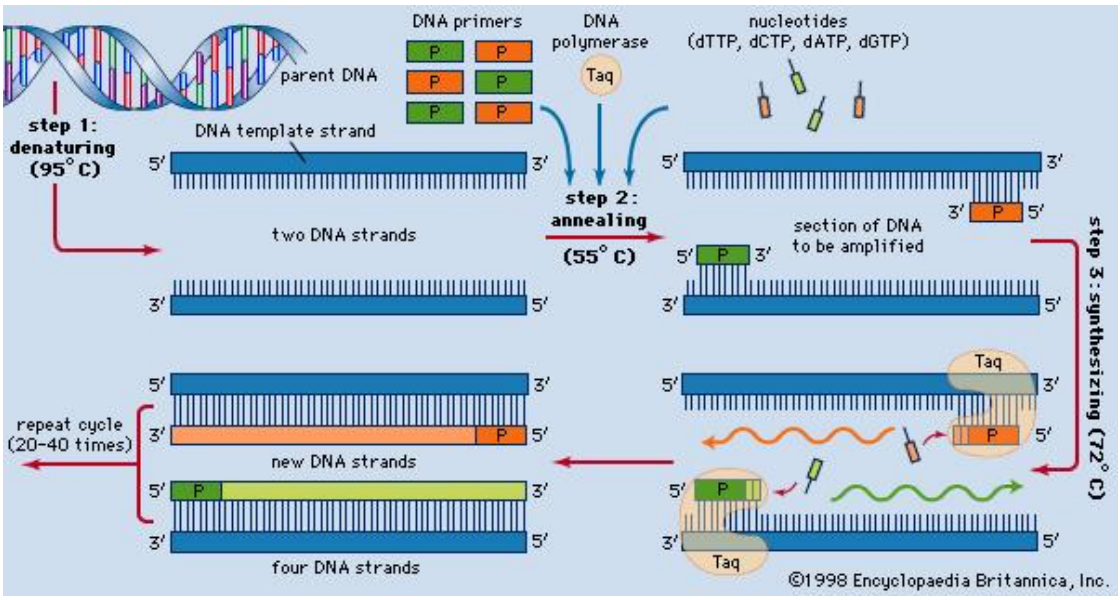

Figure 1.2: The PCR process. Image has been taken from (2).

454 pyrosequencing

454 pyrosequencing (3) was the first NGS technology to be introduced and was available commercially in 2005. This introduction revolutionised the field of metagenomics, with DNA sequencing available to many researchers around the world for the first time.

Before sequencing can take place, the DNA sample must be properly prepared (see Figure 1.3). The process begins with DNA denaturation and fragmentation. The single-stranded fragments are then attached to beads in preparation for emulsion PCR amplification which results in millions of amplicons attached to each bead in preparation for sequencing. The beads are put into wells which are then filled with helper immobilisation and enzyme beads and nucleotides repeatedly flow over the wells in the order A→T→G→C. When these nucleotides match with a nucleotide on the strands in a well (A⇔T or G⇔C), a reaction occurs and light is emitted. The intensity of the light relates to the number of nucleotides that have been attached to a particular well. The light intensities are recorded as a *flowgram* which can, in turn, be used to infer the sequences of the amplicons.

The Average read length was originally about 110 base pairs but this has since increased to over 400 and reads as long as 1000 base pairs are possible. Because the technology allows one million wells to be filled, up to one million reads are possible in a single run. Compared to other platforms, pyrosequencing is fast and produces long reads but the cost per run is relatively expensive and homopolymer length errors are possible. Much of the sequencing data analysed in this thesis was generated using 454 pyrosequencing.

Illumina Sequencing by Synthesis

Gregory et al. (4) describe the use of Illumina GAIIx sequencing to examine biodiversity. This method was developed after 454 pyrosequencing and was first available a year later in 2006. It follows the same strategy of amplifying a region of the gene and sequencing the resulting amplicons. To begin, a DNA sample is fragmented and prepared for amplification by ligation of two unique adaptors to each end of each fragment. The sample is then amplified using the chosen number of PCR cycles.

Prior to sequencing, clusters are generated using a *flow cell* (see Figure 1.4). The flow cell is coated with *oligonucleotides* which correspond to the sequences of unique adaptors ligated to the PCR amplified fragments, this allows the fragments to be bound to the flow cell. The flow cell is also coated with primers, this results in the unattached end of a fragment binding with a primer on the flow cell creating a 'bridge' with both ends of the fragment attached to the flow cell. A process called *bridge PCR amplification* follows whereby isolated clusters

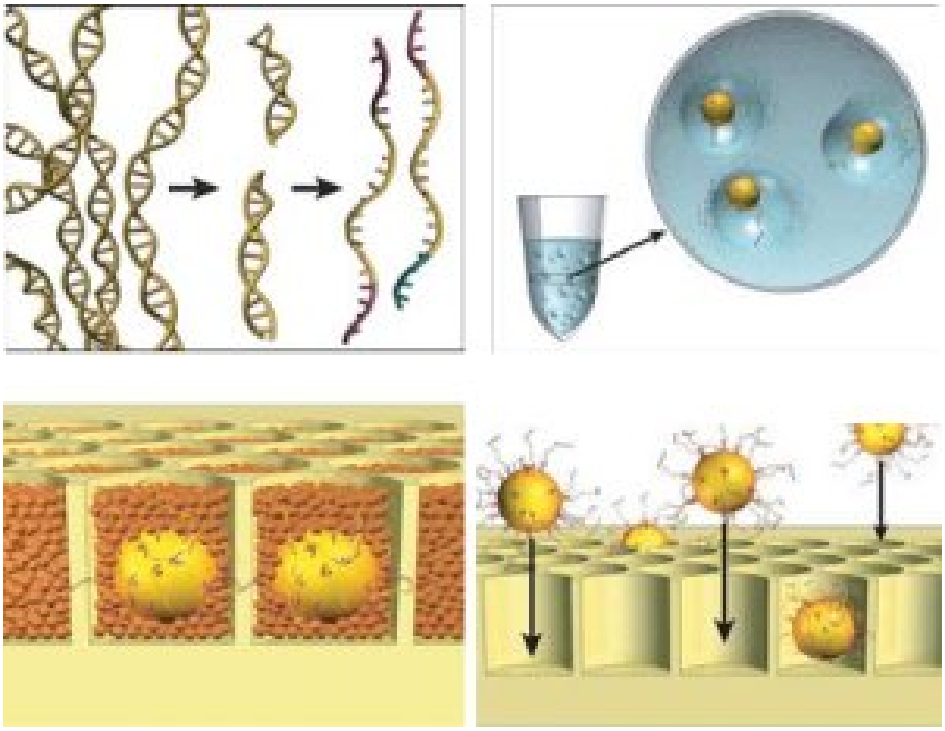

Figure 1.3: 454 pyrosequencing preparation. Top left: fragmentation and denaturation. Top right: fragments are attached to beads and amplified using emulsion PCR. Bottom right: beads are transferred to wells. Bottom left: immobilised enzymes are added to wells in preparation for pyrosequencing. Image has been taken from (3).

are amplified as the result of repeated denaturation and extension.

Sequencing by synthesis occurs with the sequential flow of nucleotides across the flow cell (see Figure 1.5). These nucleotides are fluorescent and colour coded and only one nucleotide can be bound to each strand per cycle. Excess nucleotides are washed away. Laser excitation and image capturing are used to determine which nucleotide was bound to each strand and this process is repeated until sequencing is complete.

Originally only short reads of around 50 base pairs in length were possible with Illumina sequencing, however, this later improved to around 300 base pairs. This sequencing method has the advantage of being able to produce a high number of reads per run at a low operating cost but the equipment itself is very expensive. Unlike pyrosequencing, homopolymer length errors are not an issue because only one base is read at a time but single base errors can still occur.

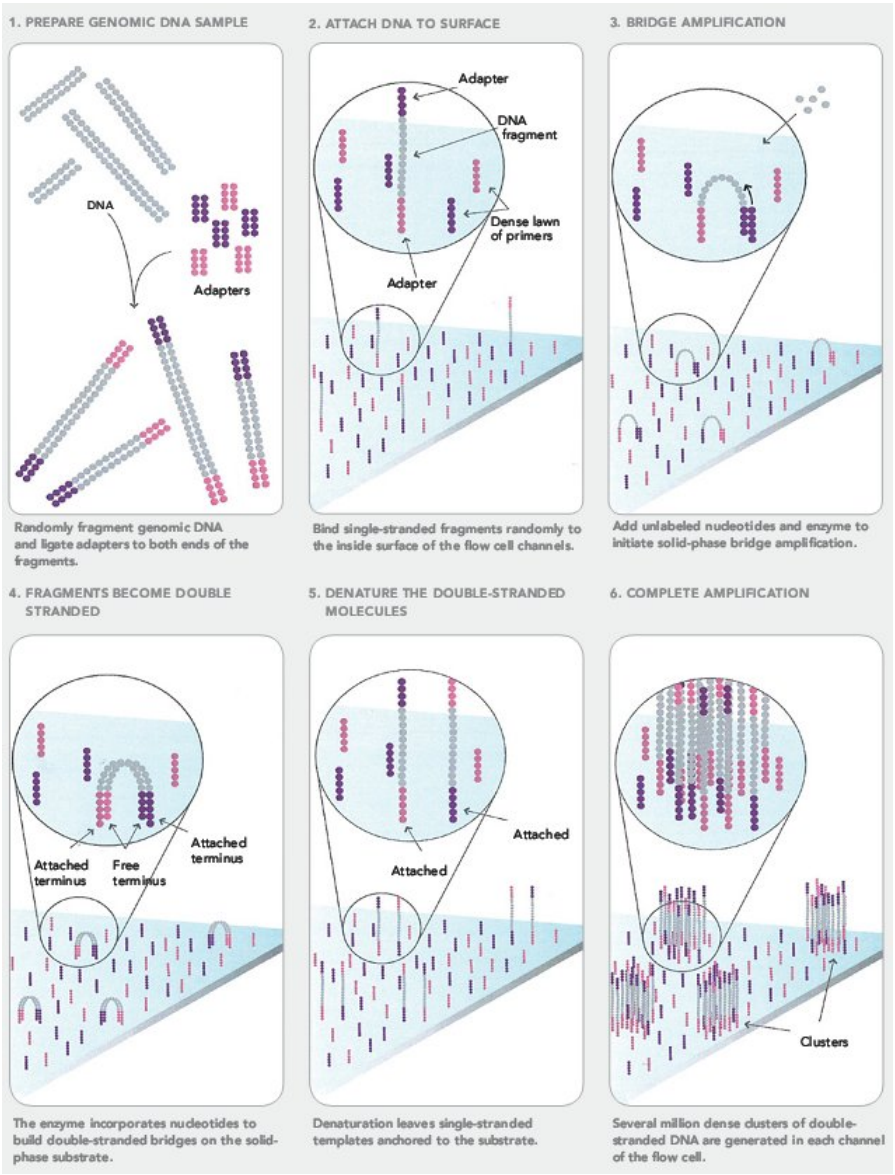

Figure 1.4: Illumina: sample preparation and bridge PCR amplification. Image has been taken from (5).

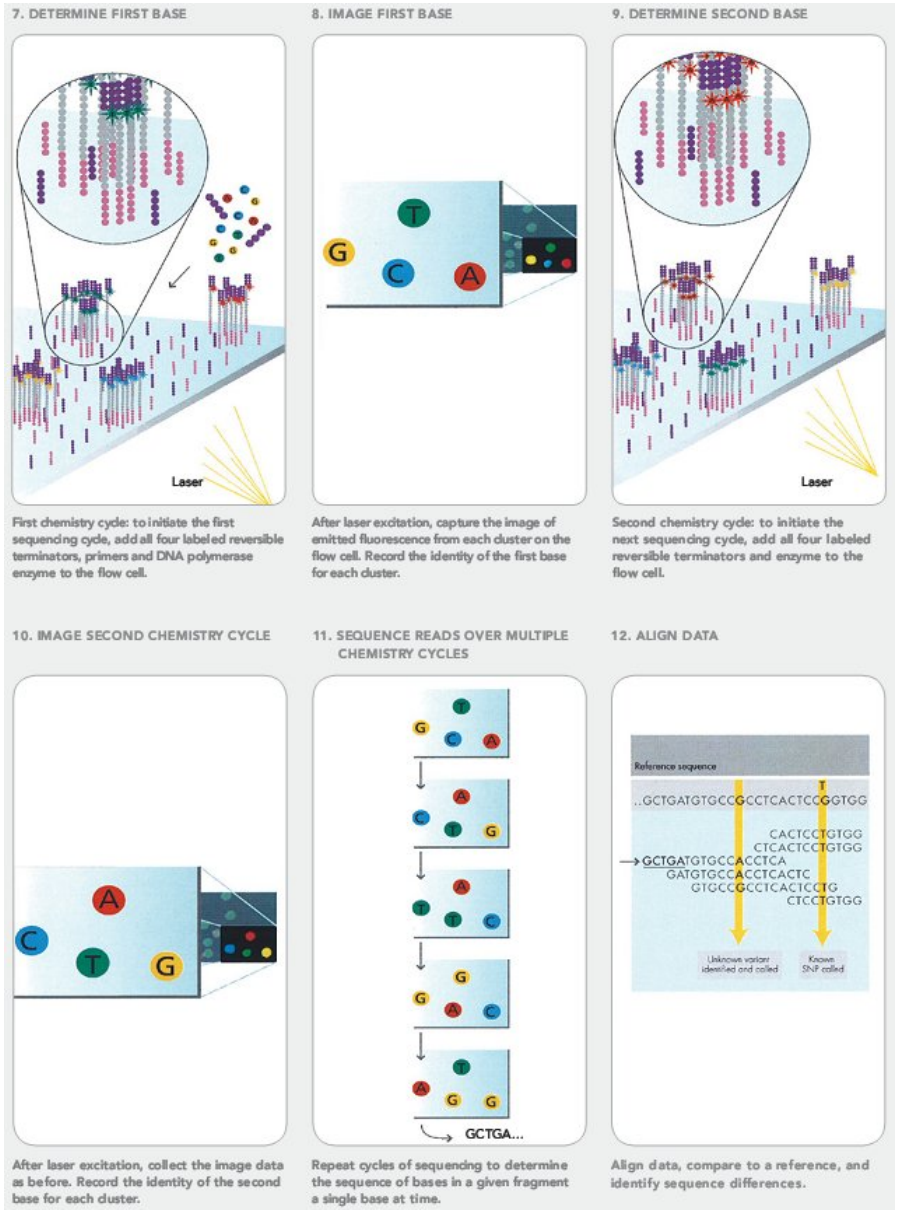

Figure 1.5: Illumina: sequencing by synthesis. Image has been taken from (5).

Ion Semiconductor Sequencing (Ion Torrent)

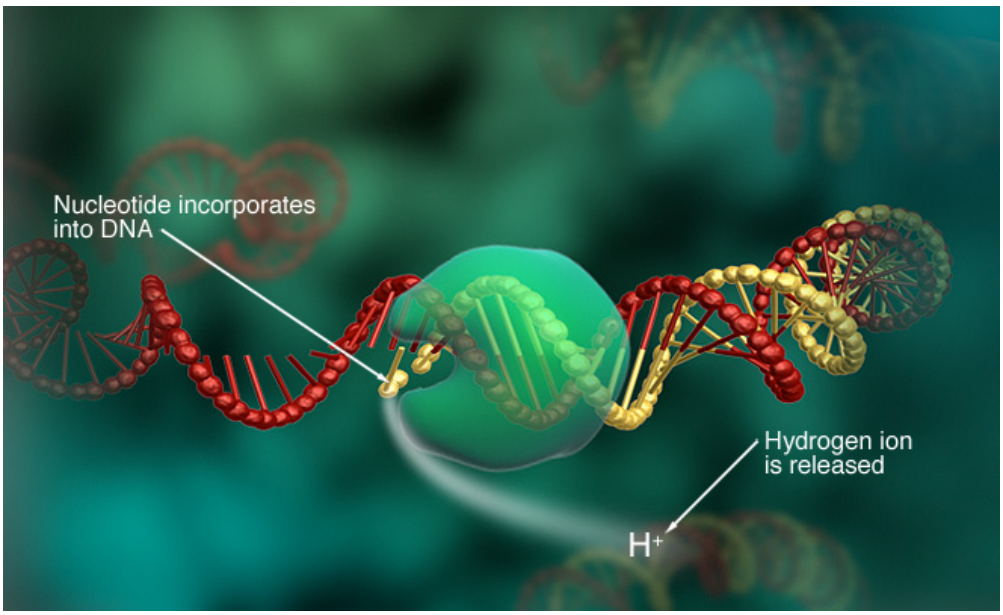

Figure 1.6: Ion Torrent sequencing. Image has been taken from (6).

Ion Torrent sequencing (7) is a method which uses semiconductor sequencing. As with the techniques used in 454 pyrosequencing, DNA is amplified using PCR and bases are sequentially flowed past the DNA strands in microwells. When a base binds with a strand a hydrogen ion (H+) is emitted which increases the acidity of the solution in the well and this change in pH is detected by an ion sensor.

Ion Torrent sequencing can produce read lengths of up to 400 base pairs. The equipment is fast and relatively inexpensive but, as the methodology is similar to that of pyrosequencing, homopolymer length errors are possible.

Sequencing by Oligonucleotide Ligation and Detection (SOLiD)

SOLiD uses fluorescently labelled di-base probes, of eight nucleotides in length, which bind to a target DNA fragment (9). These probes contain four different dyes which correspond to four unique sequences of two nucleotides in length. The probes ligate to each fragment for a chosen number of cycles, determining the eventual read length, with the first two nucleotides of the probe highlighted for each cycle. The fact that the identity of the first two nucleotides is known allows a unique sequence to be inferred from the resulting coded sequence that is detected from the coloured dyes.

This process is repeated five times using different primers, with the position of the new primer offset by one base on the first repetition and one further base on each subsequent

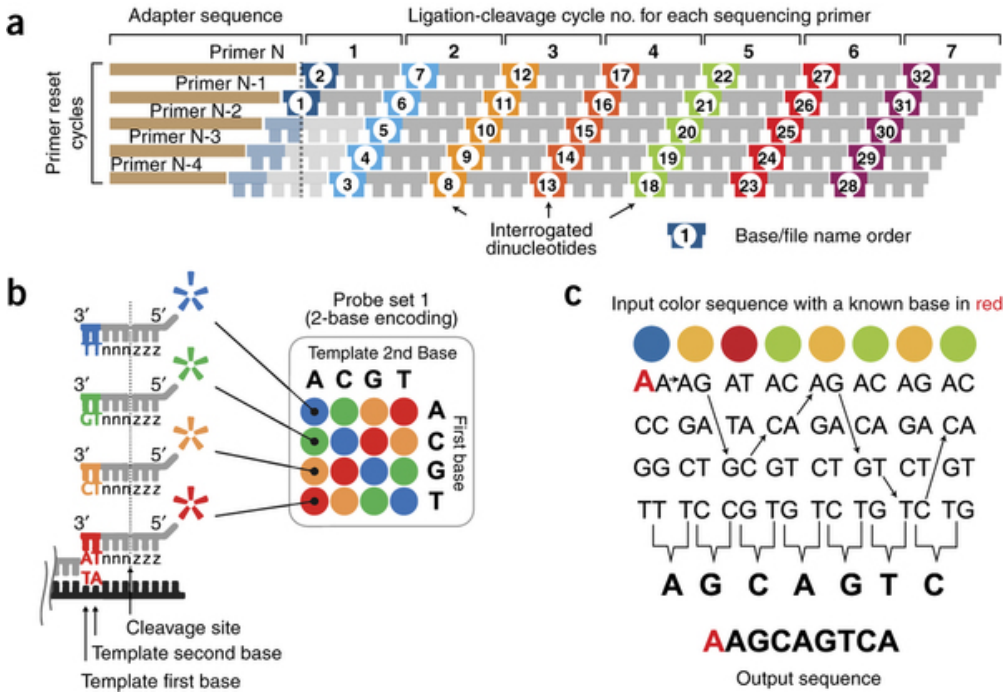

Figure 1.7: Sequencing by oligonucleotide ligation and detection (SOLiD). Image has been taken from (8).

repetition. This has the effect of interrogating each nucleotide twice in total and, therefore, results in a reduced number of errors because the same error would have to occur twice for it to go unnoticed.

The main drawback of SOLiD is that only very short reads are possible. However, due to the built-in mechanism that double checks each sequence, the error rate is relatively low compared to other NGS methods. In addition to this, homopolymer errors are impossible with SOLiD and operating costs are low.

Third Generation Sequencing Platforms

Whilst NGS technologies have been enormously beneficial, they have also created a number of problems such as sequencing noise. Limitations on throughput and read length have also fuelled further research. Most third generation sequencing technologies are still in the development phase but it is hoped that performance will be greatly improved by employing real time single molecule sequencing methods which also eliminate the requirement for PCR amplification.

Some examples of third generation sequencing platforms are *Helicos single molecule sequencing* (10), *nanopore sequencing* (11) and *single molecule real-time sequencing* (Pacific Bio) (12). The Pacific Bio platform is designed for much longer read lengths (around 10000) and the equipment runs fast but is expensive. Another disadvantage is that relatively few

reads are generated per run.

1.3 Computational Representation and Manipulation of Sequencing Data

1.3.1 Fasta Format

After sequencing has taken place, it is necessary that the information obtained is represented in an easily readable format. The format of the output from different sequencing technologies is different in each case so post sequencing processing software is supplied by the manufacturers in order to achieve this. The most commonly used representation, as mentioned earlier in this chapter, is the *fasta* format which is used extensively throughout the analysis carried out in this thesis. A fasta file contains two lines for each sequence in the dataset:

1. **Sequence name -** This line can contain any unique identifier for the sequence but it must start with the '>' character. Often it can contain extra information such as the abundance of the sequence which can be included using information returned from one of the various sequencing platforms.
2. **Sequence -** This is simply the sequence represented as a string of letters (A, C, G or T).

1.3.2 Sequence Alignment

To make sense of sequencing data in order for it to be useful, a vast array of software is available to process and analyse the data (one example of this is the noise removal software which is discussed later in this chapter). Many of these software tools make use of alignment algorithms which are used to compare sequences with each other.

When comparing the same region of a gene from multiple different organisms, it is often the case that the sequences will appear similar and exhibit only a few differences, particularly if the organisms in question are similar. It should be noted that the equivalent genes, and regions of genes, vary in length from organism to organism both within species and across species. Therefore, genetic differences can take the form not only of nucleotide substitutions but also of gaps in a sequence. These differences are caused by the genetic mutations which

drive evolution and, thus, sequence alignments can be used to better visualise the evolutionary differences between two or more organisms and hypothesise about the intermediate evolutionary steps which led to these differences.

A typical alignment algorithm will attempt to compare sequences and introduce gaps in such a way that the fewest number of differences are present at each position on the alignment. The result of this is that an aligned fasta file will have the same format as a standard fasta file with the exception that gaps (represented as the '-' character) may be included in a sequence in addition to the four nucleotides, A, C, G and T.

A *pairwise alignment* is an alignment between two sequences and a *multiple alignment* occurs between more than two sequences. Two sequences may be aligned differently within a multiple alignment than they would be in a pairwise alignment, this is because optimal pairwise alignments may contradict the goal of a multiple alignment algorithm to minimise the total number of differences contained within the full alignment. A *global alignment* has an algorithm which aims to reduce the number of differences between sequences across their entire length whereas a *local alignment* has an algorithm which searches for areas of similarity within localised regions of the input sequences.

The Needleman-Wunsch Algorithm

The Needleman-Wunsch algorithm (13) is a widely-used example of a global alignment algorithm and is best understood by following an example in which two short sequences are used as input to form a pairwise alignment.

Before the algorithm can be started, three parameters are required. These are the score for a matching nucleotide pair (m), the penalty for a mismatching nucleotide pair (mm) and the penalty for inserting a gap (g). The values of these input parameters may be varied based on different preferences for different applications of the algorithm but in the algorithm's simplest form, the parameters $m = +1$, $mm = -1$ and $g = -1$ are used. In the following example, the sequences AGTCA and AGGTCC are aligned using the aforementioned parameters.

To initiate the algorithm, a matrix is drawn with sequence A down the left hand side and sequence B along the top as shown in Table 1.2 and the cells are filled in as follows:

- A zero is inserted into the top-leftmost cell and the other cells are left empty.
- Each cell can be accessed from adjacent cells directly above, directly to the left or diagonally above and left.

		A	G	G	T	C	C
	0	← -1	← -2	← -3	← -4	← -5	← -6
A	↑ -1	↖ 1	← 0	← -1	← -2	← -3	← -4
G	↑ -2	↑ 0	↖ 2	← 1	← 0	← -1	← -2
T	↑ -3	↑ -1	↑ 1	↖ 1	↖ 2	← 1	← 0
C	↑ -4	↑ -2	↑ 0	↑ 0	↑ 1	↖ 3	← 2
A	↑ -5	↑ -3	↑ -1	↑ -1	↑ 0	↑ 2	↖ 2

Table 1.2: Needleman-Wunsch matrix for a pairwise alignment of the sequences AGTCA and AGGTCC. Green cells show one of the the best paths from the bottom right cell to the top left cell which gives one of the optimal alignments.

- If a cell is accessed from directly to the left then this represents the insertion of a gap in sequence A and the gap penalty, g, is applied.
- If a cell is accessed from directly above then this represents the insertion of a gap in sequence B and the gap penalty, g, is applied.
- If a cell is accessed from diagonally above and left then either parameter m or parameter mm is applied depending on whether the cell is a match or a mismatch.
- The best possible score for each cell is calculated, and the cell from which it was accessed is recorded. This is denoted by an arrow, as shown in Table 1.2.
- Note that sometimes more than one path will result in the same score, meaning that multiple alignments are equally optimal. For simplicity, additional alignments have not been illustrated in this example.
- When the matrix is complete, the aligned sequences can be read in reverse order from the bottom right cell by following the arrows back to the top left cell. Diagonal arrows indicate no gap, horizontal arrows indicate a gap in sequence A and vertical arrows indicate a gap in sequence B.

Reading the two sequences from the path shown by the green cells in Table 1.2 reveals that one possible optimal alignment for the two input sequences is:

```
AG-TCA
AGGTCC
```

The Needleman-Wunsch algorithm has been adapted to use more sophisticated scoring systems. For example, it may be desirable to impose heavier penalties depending on which nucleotides are involved in a mismatch. It might also be desirable to impose a heavier penalty for the first gap in a string of gaps. The algorithm is still widely used, especially in situations where alignment precision is more important than the speed of the algorithm.

The example in this section shows a simple global pairwise alignment. Other algorithms exist for local alignment (e.g. the Smith-Waterman algorithm (14)) and for multiway alignment in which algorithms must deal with increased dimensionality.

Alignment Software

A number of tools are available to generate alignments from sequences, using fasta files as input. Different types of alignment algorithm may be specified depending on the level of speed and accuracy required - a trade-off in accuracy may be required to perform a very large multiple sequence alignment. Three of the most widely used sequence alignment tools are *MAFFT* (15), *Clustal X* (16) and *muscle* (17).

1.4 Sequencing Noise

Sequencing noise is, unfortunately, a problem that is prevalent in all NGS methods. Erroneous sequences masquerade as real DNA sequences in sequencing output and, in order for meaningful analysis to take place, they must be distinguished from the genuine data. This section deals with the noise that is found in 454 pyrosequencing data.

1.4.1 Types of Noise

Noisy sequences are often created when a genuine sequence, for whatever reason, has errors applied to it. These errors can manifest themselves in two main ways which are discussed in this section and are illustrated in Figure 1.8.

Single Base Errors

Single base errors occur when one nucleotide in a sequence is substituted for an erroneous nucleotide. There are two types of single base error and the most common of these occur when a purine is substituted for another purine (A $\Leftrightarrow$ G) or a pyrimidine is substituted for another pyrimidine (C $\Leftrightarrow$ T) and is known as a *transition*. Pyramidine for purine - (A or G) $\Leftrightarrow$ (C or T) - substitutions are less common and are known as *transversions*.

Indel Errors

Insertions and deletions are categorised together as *indel* errors. These occur when a subsequence of one or more bases are added or removed from a DNA sequence.

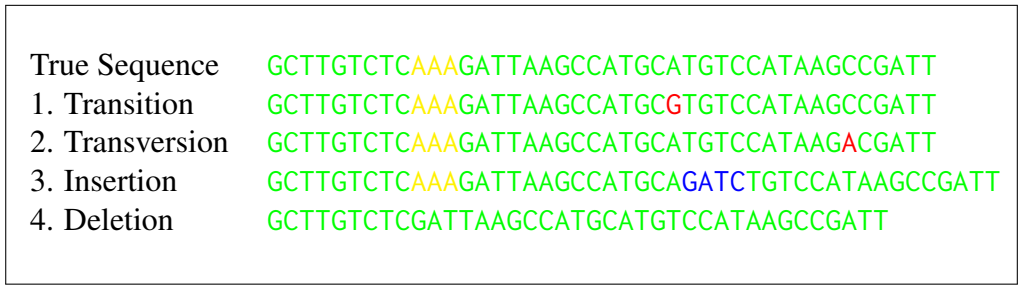

Figure 1.8: Types of sequencing noise: 1. A transition of A $\Rightarrow$ G has occurred. 2. A transversion of C $\Rightarrow$ A has occurred. 3. The sub-sequence 'GATC' has been inserted. 4. The sub-sequence 'AAA' has been deleted.

1.4.2 Sources of Noise

There are three main sources of noise in 454 pyrosequenced data. These are described by (18) and are summarised in this section.

Sequencing Errors

Noisy sequences produced in pyrosequencing are caused by *homopolymer length errors* which are a type of indel error. A *homopolymer* is part of a sequence made up of consecutive nucleotides of the same type (AA, AAA, etc.) The light intensities (continuous) do not match perfectly with the homopolymer lengths (discrete), thus the variance in the distribution of the light intensity for a given chain length is a source of sequencing noise. This variance increases with length.

PCR Single Base Errors

Single base errors may be caused during PCR when the wrong nucleotide binds to the template strand. Erroneous instances of A $\Leftrightarrow$ G and C $\Leftrightarrow$ T binding (transitions) are more likely than other errors (transversions). The probabilities of single base errors are shown in Table 4.1.

Nucleotide	A	C	G	T
A	0.9995	7.2×10^{-6}	5.1×10^{-4}	7.7×10^{-6}
C	1.1×10^{-5}	0.9996	2.1×10^{-6}	4.1×10^{-4}
G	3.5×10^{-4}	3.2×10^{-6}	0.9996	2.1×10^{-5}
T	9.0×10^{-6}	5.7×10^{-4}	1.4×10^{-5}	0.9994

Table 1.3: Probabilities of single base errors based on data from mock communities (18). Rows are the true nucleotides and columns are those observed, therefore the probabilities of true nucleotides being observed are shown on the main diagonal and the probabilities of errors are shown off the main diagonal.

Chimeras

Chimeras are PCR artefacts that are named after a monster from ancient Greek mythology. The mythological chimera was a composite of different creatures (19), specifically a lion, a goat and a snake and it lends its name to the PCR artefact because PCR chimeras are comprised of parts of different DNA molecules.

Chimeras are formed when the PCR extension step is incomplete. This results in a fragment of DNA that can act as a primer for a different sequence in another round of PCR and has the effect of forming a sequence which is really a combination of two or more different partial sequences, as shown in Figure 1.9. The proportion of chimeras present varies from dataset to dataset. Some datasets can be composed of 90% chimeric reads and this is obviously a large problem that must be addressed.

Figure 1.9: PCR chimera formation.

1.4.3 AmpliconNoise for Noise Removal

Data is initially provided as a flow file in a binary file format; this can be translated into a text file containing information, including the number of reads, the number of flows (number of nucleotides flowed across the plate), flowgram data and the read sequence. Before starting the noise removal process, the following unwanted reads are filtered out:

- Reads without the right barcode. A barcode is a short, unique sequence of nucleotides that is used to identify each read. If a read doesn't have a valid barcode then it is assumed to be corrupted and is filtered out.
- Reads without the right primer.
- Reads shorter than 360 base pairs in length.

Additionally, reads longer than 720bp are truncated at 720bp.

After filtering, the noise removal process begins by using the AmpliconNoise algorithm. This is split into two stages called PyroNoise and SeqNoise. These algorithms deal with removing noise produced during sequencing and removing point errors produced during PCR respectively.

PyroNoise

Prior to conversion into more readable formats, such as fasta format, the sequencing data generated by pyrosequencing is stored as $flowgrams$ which hold records of the light intensities that were observed. PyroNoise analyses the flowgram data in an attempt to distinguish between good reads and noisy reads.

To summarise the PyroNoise algorithm, firstly the probability is calculated that a given flowgram was generated from a sequence of nucleotides corresponding to the information from an error-free flowgram. From this, a distance metric is generated which is the negative natural logarithm of this probability, normalised by the flowgram length.

Clusters of sequences are formed based on their flowgram distances and the likelihood function of the observed data is maximised using an EM (expectation-maximisation) algorithm. For every iteration of the EM algorithm, the number of clusters decreases and the consensus sequences of the final clusters give the true denoised sequences found by PyroNoise.

SeqNoise

SeqNoise uses similar techniques to PyroNoise to eliminate single base pair PCR errors but the two algorithms are separated because it is more appropriate to use flowgram data in PyroNoise and textual sequence data in SeqNoise.

As with PyroNoise, a distance metric is calculated by taking the negative natural logarithm of the probability that a given read comes from an error-free sequence. Again, an EM algorithm

is used to maximise the likelihood function of the observed data and clusters of sequences are generated to each represent one true sequence.

1.4.4 Chimera Removal Software

Perseus

Perseus (18) is a program which generates the Chimera index for each read. This is a value greater than or equal to zero with higher values corresponding to reads that are most likely to be chimeras. Using this index, chimeras can be identified and eliminated from the data.

To begin with, the program attempts to determine the two most likely parents of the candidate read and the most likely break point. Every read is aligned with every other read of greater abundance (parent reads will be of greater abundance because they will have experienced at least one more round of PCR than the chimera). The two most likely parents are found and the break point that minimises the number of differences between the candidate read and the read created from combining the two parents is determined.

The PCR error corrected distance between these two sequences is calculated. If this value is less than 0.15 then the candidate is considered to be, potentially, a chimera. If the value is greater than 0.15 the read is classified as a good read at this point.

The next step is to find the probability of the candidate sequence evolving naturally. A three way alignment is formed, incorporating the candidate sequence and the two parent sequences. The sequence which is the common ancestor of all three sequences is found using parsimony. The following labels are used:

- A – The parent sequence that matches the candidate sequence most closely.
- B – The other parent sequence.
- C – The candidate sequence.
- D – The sequence ancestral to A, B and C.
- x – The number of differences between A and D.
- y – The number of differences between B and D.
- x_B – The number of differences between A and D on the part of the alignment matching B.

- y_A – The number of differences between B and D on the part of the alignment matching A.
- n_A – The length of the part of the alignment matching A.
- n_B – The length of the part of the alignment matching B.
- N – The total length of the alignment.

For C to have evolved naturally there must be at least x_B (out of x total) differences between A and D on the part of the alignment matching B and at least y_A (out of y total) differences between B and D on the part of the alignment matching A. The probability of this happening is

$$Pr(X_1 \geq x_B) \times Pr(X_2 \geq y_A)$$

where X_1 and X_2 are random variables such that

$$X_1 \sim Bin(x, \frac{n_B}{N})$$

and

$$X_2 \sim Bin(y, \frac{n_A}{N}).$$

This assumes an equal probability of changing for each nucleotide in the sequence.

The negative natural logarithm of this probability is the Chimera index. The lower the probability of the sequence evolving naturally, the higher the Chimera index. Perseus uses logistic regression to classify chimeras and remove them from the data.

The algorithm is only designed to remove bimeras (chimeras with two parent sequences). Trimeras and quadmeras are also possible and it is found that Perseus deals adequately with these without explicitly targeting them.

Logistic Regression

Logistic regression is a deterministic classification technique that can be used to predict whether a read is a chimera given its Chimera index, I. A logit link function is chosen so that:

$$Pr(\text{Chimera}|I) = \frac{1}{1 + \exp(-[\alpha + \beta I])} \, , I \geq 0 \, .$$

When a logistic regression is carried out on a dataset, values for α and β can be found and so

the probability of each sequence being a chimera can be calculated using the above formula. From this, a chimera index which yields a probability of 0.5 is used as the cut-off point between good sequences and chimeras.

UCHIME

UCHIME (20) utilises a different algorithm to generate a score, much like the chimera index in Perseus, which signifies the likelihood of a given sequence being chimeric.

The main step in the algorithm involves the analysis of a three-way alignment of a query sequence with its potential parent sequences. Parents are chosen either from a reference database or, as with Perseus, directly from the dataset being analysed (*de novo*). The most likely parent sequences are selected based on their similarities with opposing ends of the query sequence. The UCHIME score is based on the number of instances in which the query sequence matches one parent but differs from the other - if the sequence matches mostly the first parent at one end and the second parent at the other end then it will receive a high score.

The UCHIME algorithm makes use of the number of 'yes' and 'no' votes on each section of the query sequence. A 'yes' vote on the left hand side is defined as a position on the alignment where the query sequence matches the first parent but does not match the second parent, and vice versa for a 'no' vote. A 'yes' vote on the right hand side is recorded at all positions on the alignment where the query sequence matches the second parent but not the first. At positions on the alignment where the two parents match each other but not the query sequence, an 'abstain' vote is recorded. The uchime score is then calculated using the following equations:

$$H_L = \frac{Y_L}{\beta(N_L + n) + A_L},$$

$$H_R = \frac{Y_R}{\beta(N_R + n) + A_R}$$

and

$$H = H_L \times H_R.$$

In the above, H is the final UCHIME score, H_L and H_R are the UCHIME scores for the left and right parts of the alignment respectively, Y_L, Y_R, N_L, N_R, A_L and A_R are 'yes', 'no' and 'abstain' votes for each part of the alignment and β and n are input variables used to weight the effect of a 'no' vote.

UCHIME has been shown to have a processing speed advantage over Perseus whilst maintaining comparable levels of accuracy (20).

1.4.5 Other Noise and Chimera Removal Software

The AmpliconNoise and Perseus procedures for noise and chimera removal have been shown to work well (18) and have been integrated into some of the most used pipelines for processing sequencing data, such as QIIME (21) and Mothur (22). Other software is available either through these pipelines or as stand-alone programs to perform similar tasks which allows the user a better choice to decide on an appropriate methodology.

1.5 Analysis of Microbial and Meiofaunal Communities

It is often desirable to assess the properties relating to the makeup of a given community of organisms such as its species diversity and richness. Whilst it is relatively easy to collect the necessary data required for these analyses in larger organisms, obvious problems present themselves when dealing with communities of smaller organisms such as microbes and meiofauna due to community population size and the microscopic nature of the creatures therein. The developments in NGS technologies described in Section 1.2.3 have opened up more ways to achieve this and have allowed new analysis to take place, the scope of which was never before possible.

Strategies vary depending on the nature of the communities involved and the goals of the research to be carried out. In communities of bacteria or archaea, when selecting a gene from a sample to be sequenced, the 16S rRNA gene is often chosen because it contains a number of conserved and variable regions (labelled V1 to V9) and it is present in all species of bacteria and archaea. In order to gain data from which to analyse the diversity of a microbial community, a selection of the variable regions are amplified and then sequenced. For bacteria and archaea, the V6 region is usually incorporated into this selection because it is the most variable region and will, therefore, provide the most information about the differences between the members of the community.

For the sequencing of meiofaunal communities and individual meiofaunal organisms described in Chapter 2 and analysed in later chapters, the V1-V2 regions of the 18S nuclear small subunit (nSSU) rDNA gene were sequenced. Primers were chosen to target approximately 450bps of the gene that are known to be highly variable in meiofauna (23).

1.5.1 Operational Taxonomic Units

An *operational taxonomic unit (OTU)* refers to a group of organisms that are genetically related to each other to a specified degree of similarity. OTUs are often thought of as being equivalent to species, especially when used in reference to microorganisms for which the boundaries between species are difficult to define (as is the term *species* itself). However, the cut-off levels of similarity used for generating OTUs are generally arbitrary and are chosen to represent the relatedness deemed appropriate for the study in question. For the analysis of meiofauna sequencing data, OTUs with a 96% cut-off have been shown to most closely resemble species (24).

OTU Generation - Clustering

OTUs are generated by forming clusters of sequences based on their similarity. To initiate the process, a Jukes-Cantor (25) evolutionary distance matrix is calculated based on the number of differences between sequences when a multi-way alignment is formed. Following this, a hierarchical clustering algorithm can be applied, summarised by the following steps.

- Clusters are initialised by allocating each sequence to a separate cluster.
- The two clusters with the closest distance are combined to form one new cluster.
- The previous step is repeated until the distance between all clusters is greater than the chosen cut-off.

There are three standard variations of this algorithm which are based on how the distances between clusters are evaluated. For *complete-linkage* clustering, the inter-cluster distance is chosen to be the distance between the most distant individual members of each cluster. *Single-linkage* clustering is the opposite to this, the two closest individual members of each cluster are used. *Average-linkage* uses the mean value of the full set of pairwise distances between members of each cluster.

For OTU generation, complete-linkage is the most commonly used method because it prevents the distance between any two members of a given cluster being greater than the chosen cut-off distance. It also does not suffer from the *chaining phenomenon* in which single sequences are added to a large cluster one by one. This phenomenon is one of the disadvantages inherent in single-linkage clustering.

1.5.2 Species Richness

Species richness (or OTU richness) is the number of different species present in a community. In microbial communities, not all species will be included in a given sample so it is valuable to establish how much of the overall species richness has been uncovered. One way of doing this is *rarefaction* in which reads are randomly sampled a specified number of times for each sample size. As the sample size increases, the mean number of species or OTUs in the sampled set is recorded and plotted. If the rarefaction curve (see Figure 1.10) approaches an asymptote then most of the species in the community will be present in the sample. If the curve still has a relatively steep gradient once all reads have been sampled then this suggests that the community has been under-sampled and much species richness remains hidden.

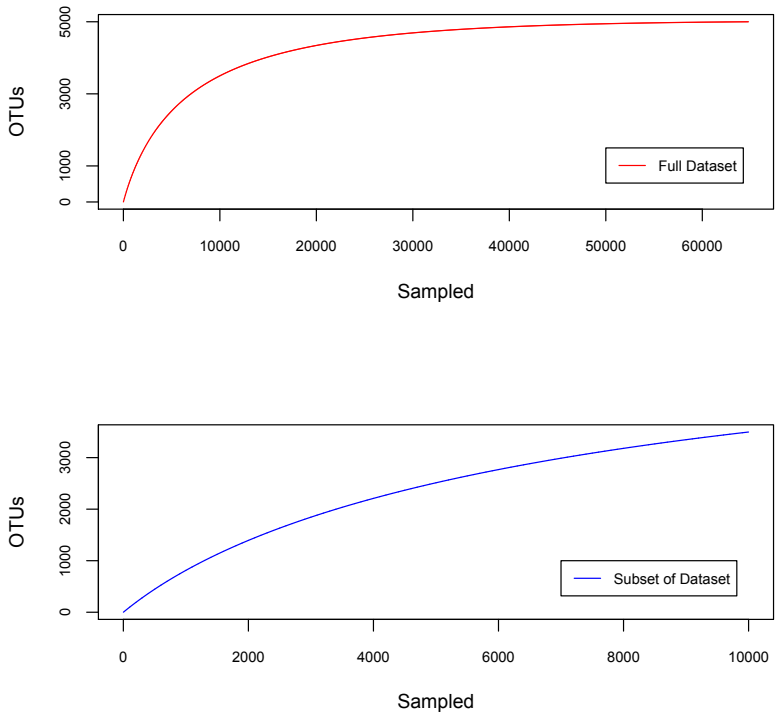

Figure 1.10: Rarefaction curves for simulated community abundance data generated from a log-normal distribution. There are approximately 63000 individuals from 5000 OTUs in the full dataset (top curve) and 10000 individuals were sampled from this to generate the bottom curve. Note that the trajectory of the bottom curve suggests a degree of under-sampling - in fact there are around 1500 OTUs missing from this sample.

Because species richness increases with sample size, it may be sensible to rarefy data when comparing multiple samples. This simply involves subsampling each sample so that they are all reduced to the same size (usually the size of the smallest sample).

Estimators of Species Richness

There are a number of different non-parametric methods of estimating the species richness based only on observed data. The *Chao1 estimator* (26) is

$$S_1 = S_{obs} + \frac{F_1^2}{2F_2}$$

where S_{obs} is the observed number of species in the sample, F_1 is the number of singleton species in the sample and F_2 is the number of doubleton species in the sample. The theory is that if rare species (singletons) are being discovered then there are likely to be yet more rare species to be found; as more of these singletons become doubletons then it becomes more likely that the majority of species have already been discovered.

The *Chao2 estimator* applies the same ideas when only occurrence data is available instead of abundance data:

$$S_2 = S_{obs} + \frac{Q_1^2}{2Q_2}$$

where Q_1 is the number of species that only occur in one sample and Q_2 is the number of species that occur in exactly two samples.

The *Jackknife estimator* (27) is calculated by

$$S_{jack} = S_{obs} + Q_1 \left(\frac{m-1}{m} \right)$$

where, again, S_{obs} is the observed number of species and Q_1 is the number of species occurring in only one sample. The variable m is the total number of samples.

The final richness estimator shown in this section is the *bootstrap estimator* (28),

$$S_{boot} = S_{obs} + \sum_{k=i}^{S_{obs}} (1 - p_i)^2$$

where p_i is the proportion of samples in which the ith species is present.

There is much debate over which estimators are most useful (29) (30) and the choice of which to employ may be dependent on the nature of the data to be analysed. For the meiofauna community data described in Chapter 2, most estimators produced similar results but the Chao1 richness estimator was chosen because it is known to function well regardless of sample size and is informative when used with data that contain many low-abundance species (30).

1.5.3 Species Diversity

Species diversity (or OTU diversity) is the relative number of species in a community. The diversity of a community is related to its richness but if the number of individuals in the community changes, and assuming that the number of species remains constant, then the diversity will change whereas the richness will not. There are a number of different diversity measures but one of the most common, and the one used in later chapters, is the *Shannon index* (31),

$$H' = -\sum_{i=1}^{S}\{p_i \ln(p_i)\}$$

where S is the total number of species and p_i is the probability of a randomly chosen individual belonging to species i. The higher the value of H', the greater the diversity of the sample is deemed to be. The Shannon diversity index is maximised when all species have equal abundance.

Alpha, Beta and Gamma Diversity

Sometimes it is useful to categorise diversity into three different types (32). *Alpha diversity* is the traditionally defined measure of diversity at one particular site and can be measured using the Shannon index.

Beta diversity is the diversity measured over a range of different sites and describes how much of the whole diversity can be seen by observing a single site. The most widely used measure of beta diversity is simply

$$\beta = (S/\bar{\alpha}) - 1$$

where S is the total number of species in all sites and $\bar{\alpha}$ is the mean species richness per site. One is subtracted from the value so that the minimum beta diversity is set to zero.

The Shannon index may also be used to calculate *gamma diversity* which is the overall diversity of all known sites.

1.5.4 Species Evenness

Species evenness (or OTU evenness) describes how balanced a community is in terms of the abundance of its composite species. For example, a community with one very dominant species will have a low evenness. Species evenness is usually measured using Pielou's

evenness (33),

$$J' = H' / \ln(S)$$

where H' is the Shannon diversity and S is the total number of species.

1.5.5 Dissimilarity Indices

It is often of interest to investigate the similarities and differences in the community content of two different samples. To do this there are several measures of dissimilarity available which are based on the abundances of different species (or OTUs, or other taxonomic ranks). The *Bray-Curtis dissimilarity index* (34) is one such measure:

$$B_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where C_{ij} is the sum of the lowest of the two abundances of individuals belonging to species that occur in both samples i and j. S_i and S_j are the total number of individuals in sample i and sample j respectively.

The Bray-Curtis dissimilarity index is also known as the *Hellinger distance*. When transformed into a similarity measure $(1 - B_{ij})$ it is equivalent to the *Sørensen similarity index*.

1.5.6 Analysis of Variance

Analysis of Variance (ANOVA) is a statistical technique which is used extensively across many different areas of research and it is used in this thesis to interpret results in Chapters 2 and 5. There are a number of different variations in the methods that are used for ANOVA but the general principle involves an investigation into the effects of explanatory variables on another variable of interest. The idea is that, typically, the variable of interest will exhibit a high degree of variance over all observations but when the dataset is compartmentalised based on the explanatory variables then the variance in each compartment will be lower. The level of reduction in the variance reveals how large of an effect, if any, each explanatory variable has on the variable of interest.

To determine whether or not a particular explanatory variable has an effect on the variable of interest, an F-test can be performed. For example, if the variable can be expressed as a number of different categories then the F test statistic is:

$$F = \frac{\text{variance between categories}}{\text{variance within categories.}}$$

The F test statistic can then be compared to the F-distribution (35) with $(c-1)$, $(n-1)$ degrees of freedom where c is the number of categories and n is the total number of observations. The p-value is taken to be the probability that a random variable generated from the F-distribution with $(c-1)$, $(n-1)$ degrees of freedom will be greater than F. A p-value lower than 0.05 is usually considered to be significant.

All standard ANOVA presented in this thesis was performed using the *lm* and *anova* functions available in **R**.

1.5.7 Species Interaction Networks

Types of Interactions

There are a number of different ways in which species can interact with each other within a community. Interactions can be labeled as one of the following types:

- **Predation** involves members of one species using members of another species as a source of food. This is clearly beneficial to the predator species and detrimental to the prey species.
- **Parasitism** is another type of interaction that is good for one species but bad for the other. The negative impacts on the host species are not as pronounced because they are not always fatal.
- **Mutualism** refers to a symbiotic relationship between two species which is mutually beneficial. One example of mutualism is a situation where a predator species feeds on the parasites of another species.
- **Commensalism** is similar to mutualism except that the relationship only benefits one species with the other unaffected. An example is a scavenger species taking advantage of food left over by a predator species.
- **Amensalism** is a relationship which is detrimental to one of the species involved and has no effect on the other. A larger organism may accidentally kill a smaller organism without receiving any benefit.
- **Competition** between two different species occurs when both species require similar resources within the community. These interactions have negative consequences for both species because there would be more resources available if one of the competing species were not present.

Food Webs

A food web is a network of all of the predator-prey interactions within a community. An organism in a food web is ranked according to its *trophic level* which describes the number of links between a predator and the environment (trophic level zero). Strictly herbivorous organisms are said to a have a trophic level of one, their direct predators have a trophic level of two and so on. Fractional trophic levels are possible if an organism predates on organisms of different trophic levels.

Interactions Based on Co-occurrence Data

A community containing many different species will be subject to a large number of interactions between these species and the cumulative effect of these interactions will influence the respective population size of each species. The quantity and variety of these interactions means that their effect on species populations is often difficult to judge but, nevertheless, attempts have been made to use co-occurrence matrices to infer interactions between species in a community.

These analyses compare community composition between different sites and suggest that some of the variation in different species' abundances is caused by interactions between species.

The methods outlined in this section may be used separately but, sometimes, a number of different strategies can be applied to the same data. In these cases, only the interactions appearing in all of the resultant networks survive to give a consensus between all of the strategies used. For example, an "ensemble network" incorporating four different measures was generated in a study by Faust et al. (36).

Correlation and Dissimilarity Matrices

One way of inferring interactions from co-occurrence data is to compute the correlation between each pair of OTUs. Two appropriate statistics for this purpose are the *Pearson correlation* (37) and the *Spearman correlation* (38). Dissimilarity indices such as the *Bray-Curtis index*, also described in Section 1.5.5, and the *Kullback-Leibler divergence* (39) can also be used to infer interactions based on how different samples are with respect to their OTU composition.

More detail about these indices and their usage in the context of co-occurrence data analysis

is provided in Section 5.4.4. In each case, the result of the analysis will be a square matrix containing values to show how correlated or dissimilar each pair of OTUs are to each other.

SparCC

Some of the more basic methods of interaction inference are somewhat unreliable because the relative abundances found from sequencing data do not correspond directly to the true number of organisms belonging to each species. These abundances can be affected by variance in the number of copies of a gene present in different species. SparCC is another approach used to infer correlation values from co-occurrence data which takes this into account and has been shown to produce good results on simulated data (40).

Results show (40) that the aforementioned compositional ambiguity is driven by OTU diversity within communities and the SparCC algorithm makes use of these findings to formulate a matrix of correlations which can be used to predict an interaction network. The true network of interactions is assumed to be *sparse*, meaning that most potential interactions between OTUs do not exist or are negligible. The SparCC approach is described in more technical detail in Section 5.4.2.

Local Similarity Analysis

Local Similarity Analysis (LSA) is another technique which can be used to investigate relationships between pairs of OTUs and was first introduced in 2006 by Ruan et al. (41). LSA can be used to detect similarities between OTUs in different samples but it is applied when the samples are part of a study using time series. For this reason it was not used for any of the research presented in this thesis but a brief overview follows in this section.

LSA requires time series data that has undergone a normal transformation (42). Normally transformed abundance data is observed for two OTUs over n time intervals to give two time series, $O_{11}, O_{12} \dots O_{1n}$ and $O_{21}, O_{22} \dots O_{2n}$. An integer value, D, is chosen to specify the maximum distance between time series points that an interaction can take place. A positive score matrix, P, and a negative score matrix, N, both with dimensions of $n \times n$, are then calculated using the following algorithm taken from (41):

- For $i, j = 1, \dots n$:
  $P_{0,i} = P_{j,0} = 0$ and $N_{0,i} = N_{j,0} = 0$.
- For $i, j = 1, \dots n$ with $|i - j| \leq D$:
  $P_{i+1,j+1} = \max[0, P_{i,j} + O_{1,i+1} \times O_{2,j+1}]$ and
  $N_{i+1,j+1} = \max[0, N_{i,j} - O_{1,i+1} \times O_{2,j+1}]$.

- $P(O_1, O_2) = \max[P_{i,j}]$ for $1 \leq i, j \leq n$ and
  $N(O_1, O_2) = \max[N_{i,j}]$ for $1 \leq i, j \leq n$.

- $\text{MaxScore}(O_1, O_2) = \max[P(O_1, O_2), N(O_1, O_2)]$ and
  $\text{Flag}(O_1, O_2) = \text{sign}[P(O_1, O_2) - N(O_1, O_2)]$.

The local similarity score of the two time series, $LS(O_1, O_2)$ can be calculated using the formula,

$$LS(O_1, O_2) = \frac{\text{MaxScore}(O_1, O_2)}{n}.$$

Whether $LS(O_1, O_2)$ is for positive or negative correlation between the two series is found from the sign returned by $\text{Flag}(O_1, O_2)$.

1.5.8 Ecological Models – Neutral Theory versus Niche Theory

Hubbell's *unified neutral theory of biodiversity and biogeography* (or just "neutral theory") (43) states that every species will have the same chance of success per capita as every other species that shares the same trophic level on a food web. The theory claims that random events are the dominant force which determine which species will make up a community and that competitive advantages of certain species are negligible. It follows that, in this model, the level of diversity at any location is driven entirely by chance.

In contrast, *niche theory* (44) states that species will tend to occupy environments (niches) that they are most suited to. It also states that competition between two trophically similar species for the same niche can result in one of these species being driven away if it is significantly worse at adapting to that niche. Niche modelling can be used to predict the presence of certain species in a location before they are observed there based on environmental factors. For example, if one site is sufficiently similar to another site (i.e. it provides the same niche) then it is likely that the community composition of the two sites will be similar.

In practice, both theories can be applied in certain circumstances. Sometimes, within the same community, a neutral model may be appropriate for some species but a niche model will be a better representation for others. In simple terms, some species are more sensitive to their environment than others.

1.6 Thesis Overview

Chapter 2 describes two experiments that were carried out between 2007 and 2008. The first of these was a study of meiofauna communities in sand sediment samples collected from various sites in Europe and one in Africa (Gambia). The aim of this study was to investigate the distribution and diversity of the various meiofauna phyla in the collected samples. The samples were pyrosequenced (Section 1.2.3) and the resulting sequencing data were processed for noise removal (Section 1.4). OTUs were generated and many of the methods outlined in Section 1.5 were utilised to reveal valuable new information about the distribution and diversity of meiofauna.

The second experiment described in Chapter 2 involves the analysis of individual nematode samples and pooled nematode samples, with the component nematodes selected based on their phylogenetic relatedness to each other. Again, samples were pyrosequenced and processed for noise removal (Section 1.2.3 and Section 1.4). The results from the Perseus chimera detection software were used to investigate drivers of chimera formation, answering questions about the effects of sample composition and the nucleotide diversity at various regions of genes.

Chapter 3 builds on the knowledge of how PCR works in practice (Section 1.2.3) and, in particular, how PCR chimeras are formed (Section 1.4.2) in order to design algorithms to simulate chimera formation in PCR. The implementation of these algorithms are tested and parameters are calibrated using the results relating to chimera formation found in Chapter 2. A good PCR algorithm is an important development because it can be used as part of the generation of *in silico* community datasets which allow fast and inexpensive analysis to be performed on, for example, the appraisal of chimera detection software.

The topic of Chapter 4 is the generation, using the software developed in Chapter 3, and subsequent analysis of *in silico* microbial communities. Because the full composition of such a dataset is known (i.e. there is no hidden diversity) then a clear picture can be drawn from its analysis using the techniques outlined in Section 1.5. It can be seen how well these techniques really perform when attempting to analyse microbial communities. The true impact of noise can also be illustrated because, in an artificial dataset, all noisy reads will be flagged as such and so there is less room for ambiguity.

Although the experiments presented in Chapter 2 were not designed for this purpose, Chapter 5 demonstrates their versatility by using the resultant data for the generation of interaction networks, including food webs of predator-prey interactions (Section 1.5.7). Co-occurrence

data from the meiofauna community analysis is used to infer interactions between species. It is also hypothesised that foreign DNA which was sequenced during the individual nematode experiments is part of the main individual nematode's diet. This is intriguing because it offers a new method for inferring predator-prey interactions between species that are too small for this to easily be achieved by observation.

Chapter 2

DNA Sequencing Experiments on Meiofauna

2.1 Introduction

2.1.1 Credit for Experiments and Analysis

Much of the analyses carried out in Chapters 3–5 are based on two pre-existing studies on marine benthic communities of meiofauna (45) and individual meiofaunal organisms (46). This chapter describes both of these experiments and presents, in detail, the analysis carried out on the resultant data.

All of the analyses presented in this chapter were either carried out by Ben Nichols as part of the collaborative effort for the publication of the two above cited articles or have been repeated independently by Ben Nichols for inclusion in this thesis. All figures presented in this chapter have been produced by Ben Nichols from the available experimental data except where explicitly stated. Data collection and laboratory work were carried out by other authors of the above cited articles and the methods have been presented in this chapter to enable a complete understanding of the analyses that were undertaken.

2.1.2 Terminology: The Marine Benthos, Meiofauna and Protists

The *marine benthos* is the name given to the community of organisms that dwell in the sandy sediment on the sea bed. A subgroup of these organisms are known as *meiofauna* which are small invertebrates categorised by their size. Because of the variability of the within-species sizes of these organisms, the definition of meiofauna is not precise. However, an approximate guide is to include organisms which are too large to fit through a 45μm mesh but small

enough to fit through a 1mm mesh. Abundant meiofauna phyla include *nematodes*, *platyhelminthes* and *arthropods*.

Protists are a large and genetically diverse group of eukaryotic unicellular organisms (or multicellular organisms without specialised tissues) which do not have much in common with each other apart from their simple structure. Some metazoan (animal) protists, also known as *protozoa*, can be found in the marine benthos and were analysed in conjunction with the meiofauna for comparison. Examples of protist phyla are *alveolata*, *cercozoa*, *rhizarea* and *stramenopiles*.

2.2 Experiment 1: Metagenetic Analysis of the Distribution and Diversity of Marine Benthic Meiofauna

2.2.1 Introduction

Macroecology is the study of communities of organisms and the relationships of these organisms with their environment over large areas. Macroecological studies attempt to explain why communities differ in composition based on their location and how much variation is attributable to the changes in environment from location to location.

This study focuses on investigating the macroecology of meiofauna living in various marine benthic sites across Europe and Africa using next-generation sequencing techniques. Little is known about the macroecology of meiofauna when compared to that of larger organisms. The obvious reasons for this are that meiofauna are harder to observe and are much more diverse than larger animals. As a result, there is much debate about the nature of the distribution patterns of meiofauna and other small organisms - an example of this being the confusion over the wide distribution of marine meiofauna, despite the fact that these organisms don't usually have a planktonic larval stage (47).

The availability of next-generation sequencing has changed the way microbial communities are analysed (48) and, more recently, these strategies have been applied to communities of small eukaryotic organisms (49) (50) (51) (52). The studies outlined in this chapter aimed to apply these methods towards the marine benthos meiofauna which is comprised of many understudied species-rich phyla and, previously, had received little analysis in these regards.

2.2.2 Materials and Methods

Sample Collection

Samples of sandy sediment were collected at 23 locations - shown in Table 2.1 - around the UK (16 sites), France (2 sites), Spain (2 sites), Portugal (2 sites) and Gambia (1 site) during the summers of 2007 and 2008. The samples were obtained from the low-tide mark using a standard corer methodology (53) - 44mm diameter $\times$ 100mm cores were used to collect three samples at each site, approximately 10m apart from each other. The locations of the sampling sites are shown in Table 2.1 and Figure 2.1.

In addition to this a further core was taken for sediment analysis using a Malvern Mastersizer 2000 as in (49). Environmental data was also obtained - seawater salinity from DEFRA (54) and seawater surface temperature from NOAA (55).

Sampling Site	Abbreviation	Country	Latitude	Longitude
Prestwick	PWK	UK	55° 30′ 28.86″ N	04° 37′ 29.34″ W
Littlehampton	LH	UK	50° 48′ 07.56″ N	00° 32′ 23.10″ W
Mersey Egremont	EGR	UK	54° 29′ 11.28″ N	03° 36′ 17.58″ W
Moggs Eye	MEye	UK	54° 54′ 18.54″ N	01° 21′ 14.22″ W
Skye Staffin	SkyeStaf	UK	57° 38′ 09.24″ N	06° 13′ 44.52″ W
Dunnet Bay	DBay	UK	58° 36′ 52.08″ N	03° 21′ 02.34″ W
Seaham	Seah	UK	54° 51′ 16.86″ N	01° 20′ 40.02″ W
Exe	Exe	UK	50° 36′ 27.90″ N	03° 30′ 29.28″ W
Harwich	HW	UK	51° 56′ 13.50″ N	01° 17′ 25.68″ E
Sheerness	Sheer	UK	51° 26′ 24.66″ N	00° 45′ 50.64″ E
Porthtowan	Porthw	UK	50° 28′ 01.44″ N	05° 02′ 08.88″ W
Newborough	Newb	UK	53° 08′ 36.78″ N	04° 24′ 22.98″ W
Firth of Forth	FirthF	UK	55° 52′ 22.32″ N	02° 04′ 53.52″ W
Fraserburgh	Fraser	UK	57° 40′ 35.64″ N	01° 59′ 52.38″ W
Freshwater West	FreshW	UK	51° 39′ 27.12″ N	05° 03′ 50.46″ W
Silecroft	Silecr	UK	54° 12′ 57.66″ N	03° 21′ 17.10″ W
Praia Limpa	PrLimpa	Portugal	37° 05′ 27.48″ N	08° 27′ 19.20″ W
Vila Nova de Milfontes	VNM	Portugal	37° 43′ 26.70″ N	08° 47′ 33.36″ W
Mera	Mera	Spain	43° 22′ 41.88″ N	08° 20′ 16.50″ W
Sada	Sada	Spain	43° 20′ 34.02″ N	08° 14′ 22.26″ W
Cap Ferret	CapFer	France	44° 20′ 40.32″ N	01° 16′ 33.90″ W
St. Jean	StJean	France	43° 23′ 40.08″ N	01° 39′ 37.02″ W
Gambia	Gamb	Gambia	13° 28′ 08.52″ N	16° 39′ 51.72″ W

Table 2.1: Abbreviations and geographical information for the 23 sampling sites.

The samples were stored and preserved in 500ml storage pots with 300ml of DESS (20% DMSO and 0.25M disodium EDTA, saturated with NaCl, pH 8.0) (56). The whole core

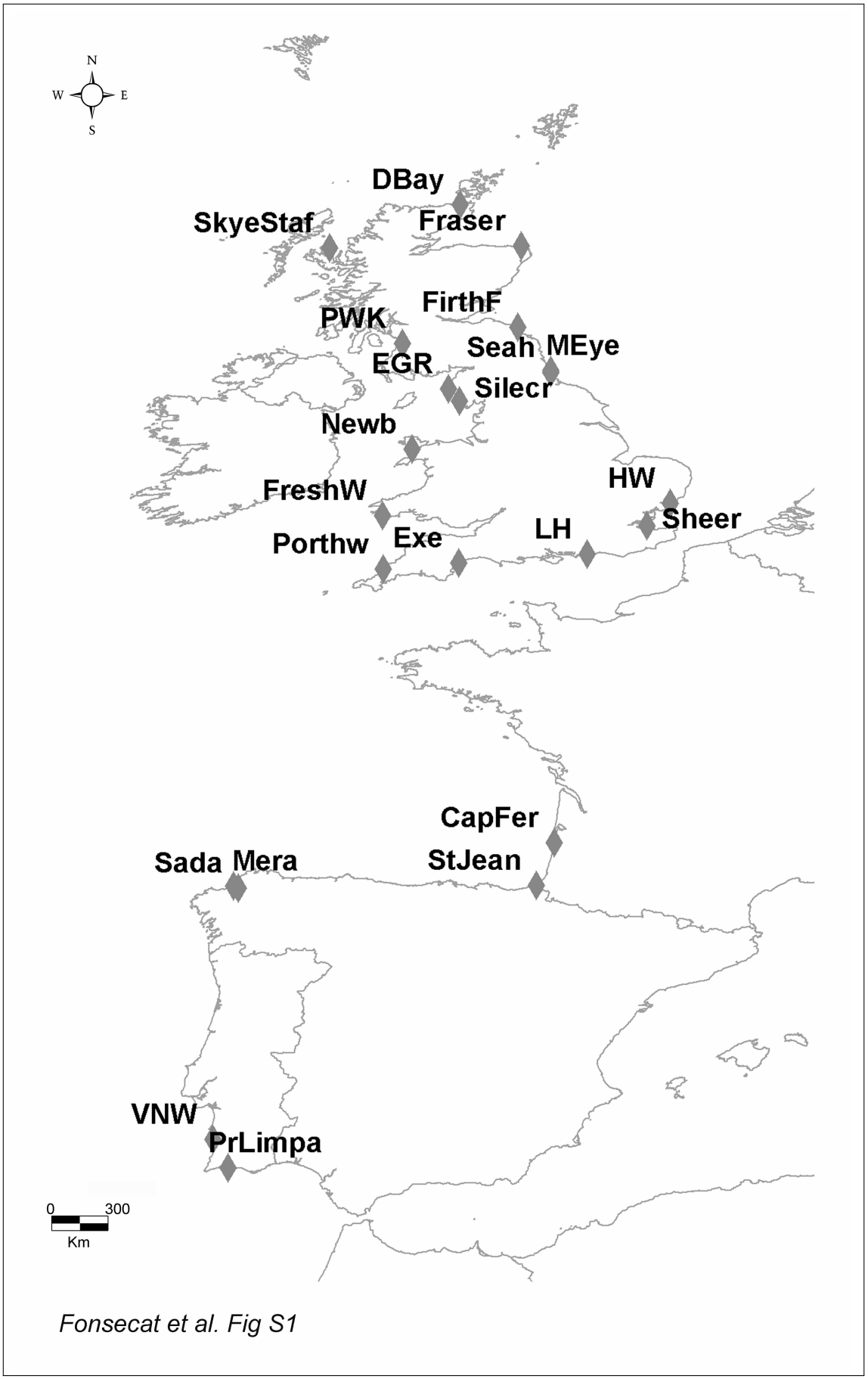

Figure 2.1: Map of the 22 European sampling sites. The 23rd sampling site is located in Gambia (not shown). Figure taken from (45).

from each sample site was used to administer meiofaunal size faction and DNA extraction (49) (23).

Primer Design and PCR Strategy

Primers were chosen to anneal to regions of the 18S rDNA gene that are highly conserved in meiofauna and are either side of a highly variable region which is ideal for the selection of operational taxonomic units (OTUs) (23). These primers were the forward primer:

SSU_FO4 (5'-GCTTGTCTCAAAGATTAAGCC-3')

and the reverse primer:

SSU_R22 (5'-GCCTGCTGCCTTCCTTGGA-3')

which amplified approximately 450bp of the V1-V2 regions of the nuclear small subunit rDNA (18S rDNA).

Fusion primers, PCR amplification and 454 sequencing were carried out using the procedures outlined in (23) and (49).

Data Analysis and Generation of OTUs

Four half-plates of 454 Roche GSFLX pyrosequencing generated sequences which were then processed using AmpliconNoise for denoising (18). Short sequences (those with fewer than 199 bp) and singletons were removed, resulting in an average sequence length of 200-220 bp. Chimeras, like any erroneous extra sequence, are known to artificially inflate diversity levels and so were identified and removed using Perseus (18).

After noise removal and chimera checking was complete, a distance matrix was calculated for all sequences to show how similar each sequence is to every other sequence. Two different sets of OTUs were generated using a complete linkage clustering algorithm, a "farthest neighbour" clustering method which involves measuring the distance between the most distant members in each cluster and grouping clusters together based on this distance.

In the first set, OTUs were represented by clusters of sequences with at least 99% similarity within each cluster and in the second set, 96% was chosen to be the cut-off. The 99% cut-off was chosen to investigate the distribution of intra-species genotypic diversity. The

96% cut-off was chosen because the AmpliconNoise analysis of a reference nematode community (24) shows that this level of similarity closely resembles actual taxonomic species richness. Therefore, the 99% OTU clustering can be thought of as the *distribution metric* and the 96% OTU clustering can be thought of as the *richness metric*.

Megablast was used on the GenBank/EMBL/DDBJ nucleotide database for taxonomic assignment. The OCTUPUS annotation and parsing toolkit (49) was used for OTU annotation and this was restricted to matches of 90% or better.

Diversity and Community Analysis

The fewest number of reads, prior to noise-removal and clustering, generated for any of the 23 sites was 9490. To standardise the data, this number of sequences were randomly selected from each site so that they each contained 9490 sequences (218270 in total).

Site-specific rarefaction curves were created using the DiversityEstimates software available in AmpliconNoise and phylum specific rarefaction curves were generated through EstimateS 8.2.0 (57) which uses a variety of different richness estimators. The *Chao1* estimator was chosen because it is not greatly affected by sample size and is most informative when used on datasets which are skewed towards the low-abundance classes (30) and is therefore applicable for datasets involving an unevenly distributed selection of microorganisms which exhibit both of these properties. This analysis was repeated using the *specaccum* (species accumulation) function in the *Vegan* package in **R**.

Cluster dendograms and multidimensional scaling (MDS) with 50 random starts were generated using PRIMER 6 (58). This required the computation of Sørensen's similarity coefficient among samples using a presence/absence similarity matrix. This analysis was repeated using the *hclust* function in **R**.

PRIMER 6 was also used to perform a similarity profile test ('SIMPROF') permutation test which is designed to test whether similarities observed in the data are of greater or smaller magnitude than those expected by chance. A permutational multivariate analysis of variance ('PERMANOVA') was also performed to test for significant differences in the composition of the samples obtained from different sites. These analyses were based on 1000 different permutations of Sørensen's similarity coefficient calculated using untransformed presence/absence data from all sites.

To test if there was a relationship between geographical distance (minimum coastal dispersal

distance between sites) and the composition of the samples a Mantel-type test ('RELATE'), using Primer, was carried out on two distance matrices - the first was the distance matrix of geographical distances and the second was the community composition (presence/absence data).

Similar RELATE tests were carried out on euclidean distance matrices calculated from the recorded environmental variables - seawater salinity, seawater surface temperature and sediment grain size in order to determine the effect of these variables on community composition. In order to reduce the effect of false positives, sequential Bonferroni corrections were applied where appropriate because these are considered to be more sensitive to false positives than standard Bonferroni corrections (59).

To find out the most useful geographic and/or environmental parameters for describing patterns occurring within each phylum, the *adonis* function in the *Vegan* package in **R** was used. This function performs a partition multivariate analysis of variance which partitions distance matrices among sources of variation and performs permutation tests to determine the significance of the partitions, in this case Bray-Curtis distances were calculated for each phylum against the environmental and geographical parameters (sea water temperature, sea water salinity, sediment grain size and latitude). The permutation tests work by generating 999 random permutations of the observed data and performing ANOVA on each of these. The F-statistics returned from these tests are compared with the F-statistics returned from an ANOVA test on the true data to calculate the p-values which determine the significance levels.

It has been shown that Hubbell's Neutral Theory of Biodiversity (43) can be approximated as a hierarchical Dirichlet process (60). In order to investigate the appropriateness of a neutral model when applied either to individual phyla or to all phyla present in this study, a hierarchical Dirichlet process was fitted to the community data using a Bayesian strategy. From the fitted model, neutral metacommunities were generated (one of each phylum and one for the all phyla combined) and the likelihoods of the abundances in these metacommunities were compared with the likelihoods of those in the corresponding observed datasets. The proportion of these likelihoods that exceeded the observed value was recorded as a pseudo p-value that the data followed a neutral model. This analysis could be used to investigate localised neutrality or neutrality across all sites.

ANOVA in **R** was used to assess which explanatory variables the neutrality, or lack of, exhibited by particular phyla could be attributed to. The variables that were investigated were phyla richness and whether the phyla was meiofaunal or protist.

2.2.3 Results

Sequence Data and Sampling Efficiency

After denoising and chimera removal the total number of reads generated from all sampling sites was reduced from 877423 to 694802. Figures 2.2 and 2.3 are rarefaction curves with an OTU cut-off of 96% showing that sequencing effort was incomplete for most samples with a significant proportion of the existing diversity at species level unidentified.

Figures 2.4 and 2.5 are similar rarefaction curves, with a 99% OTU cut-off, showing that a significant proportion of within species diversity remains unidentified.

Community Diversity, Composition and Richness

The proportion of 99% OTUs in each sampling site that were shared with at least one other sampling site and, by association, the proportion of unique OTUs are shown in Figure 2.6. It can be seen that the outgroup reference site in Gambia has the highest proportion of unique OTUs (74%), suggesting that geographical distance plays a part in the composition of meiofauna communities. Cap Ferret (France) and Sheerness (UK) had the next highest proportion of unique OTUs with 60% and 53% respectively.

Figure 2.7 presents the same information for 96% OTUs. The sites are generally distributed in the same way, with those that had a higher proportion of unique 99% OTUs also having a higher proportion of unique 96% OTUs and those that had a lower proportion of unique 99% OTUs also having a lower proportion of unique 96% OTUs. In particular, the three sites with the most unique OTUs - Gambia, Cap Ferret and Sheerness - are unchanged. One exception to this was Seaham which was ranked fifth in terms of its unique 96% OTU proportion but only fourteenth at the 99% level.

At all sites, the proportion of unique 96% OTUs was lower than the proportion of unique 99% OTUs and overall there was a lower proportion of unique 96% OTUs than unique 99% OTUs (27% versus 39%). This suggests that there is more variation in community composition at the species level than the phylum level and confirms that there are more rare species than rare phyla. This is clearly the case because species belonging to rare phyla will be rarer still.

The bulk of all unique and shared OTUs were made up by nematodes and platyhelminthes

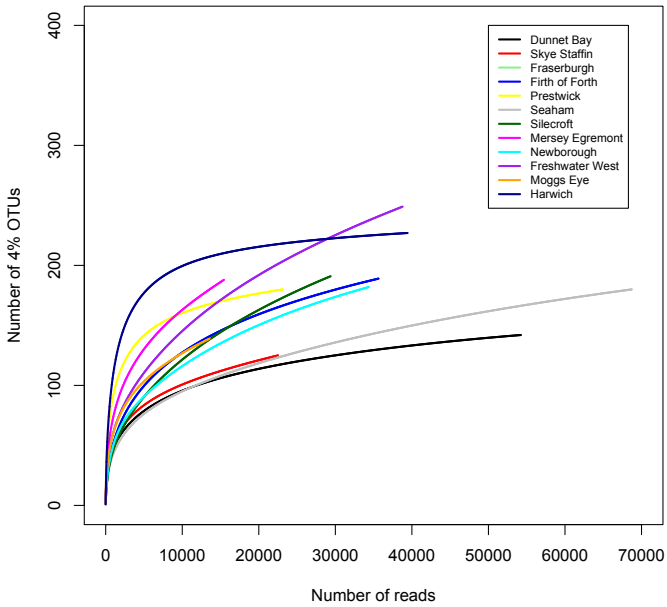

Figure 2.2: Rarefaction curves for the first 12 sampling sites using 96% OTUs. Denoised reads were clustered into OTU groups of 96% or greater similarity. Subsamples of increasing size were taken and the number of unique OTUs in each subsample was plotted.

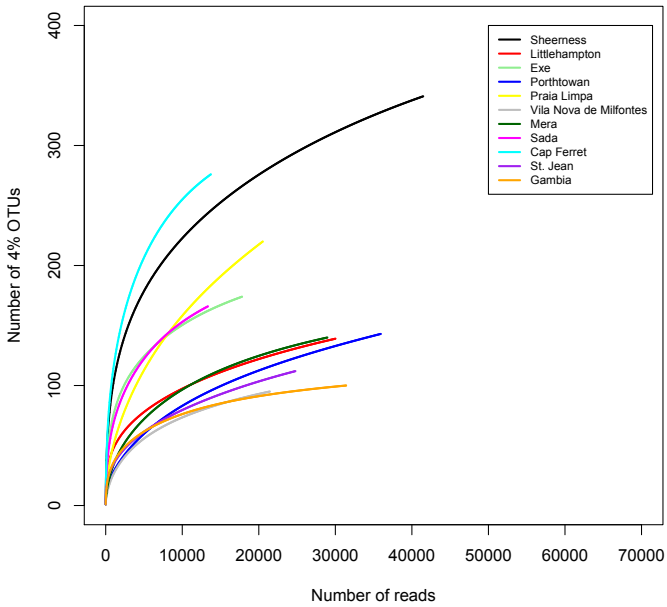

Figure 2.3: Rarefaction curves for the final 11 sampling sites using 96% OTUs. Denoised reads were clustered into OTU groups of 96% or greater similarity. Subsamples of increasing size were taken and the number of unique OTUs in each subsample was plotted.

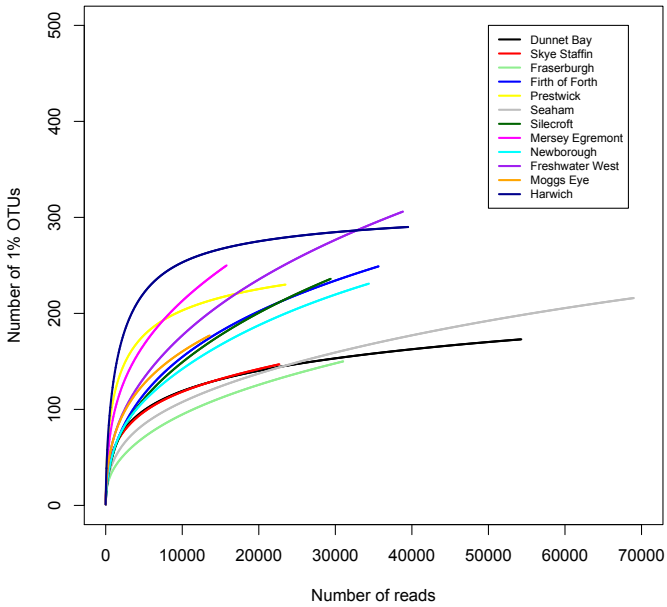

Figure 2.4: Rarefaction curves for the first 12 sampling sites using 99% OTUs. Denoised reads were clustered into OTU groups of 99% or greater similarity. Subsamples of increasing size were taken and the number of unique OTUs in each subsample was plotted.

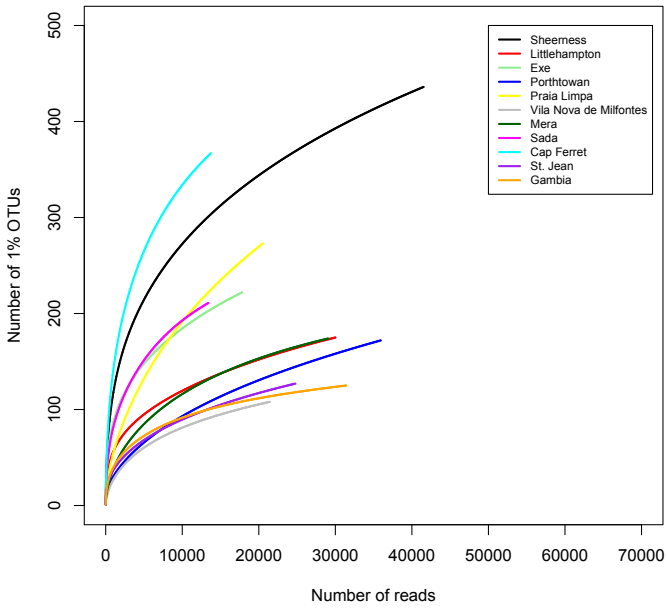

Figure 2.5: Rarefaction curves for the final 11 sampling sites using 99% OTUs. Denoised reads were clustered into OTU groups of 99% or greater similarity. Subsamples of increasing size were taken and the number of unique OTUs in each subsample was plotted.

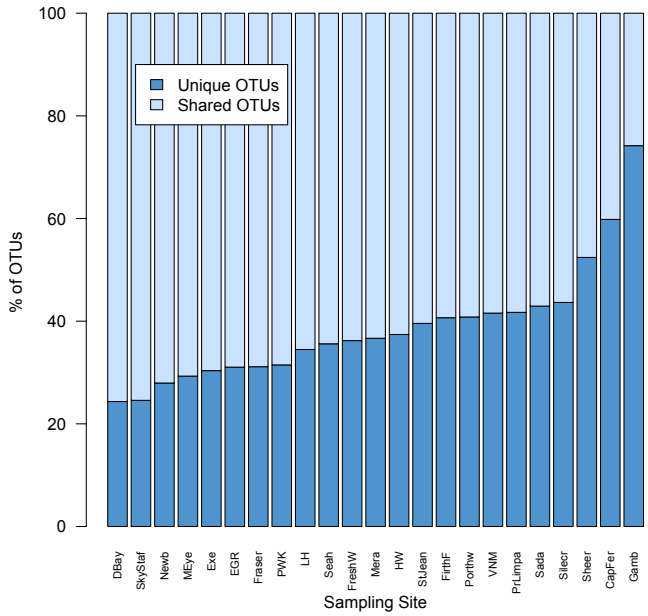

Figure 2.6: Unique and shared OTUs (99%) at each sampling site. A shared OTU is defined as an OTU that appears at more than one sampling site.

with nematodes contributing more to the unique OTUs and platyhelminthes contributing more to the shared OTUs, as can be seen in Figure 2.8.

The main meiofauna phyla are investigated in terms of their richness and distribution across sites in Figures 2.9 and 2.10. The data were clustered into 96% OTUs which were each assigned to the correct phylum. Figures 2.9 and 2.10 suggests that the distribution of phyla in continental European sites is more heterogeneous than those in the UK which are more dominated by the abundant meiofauna such as nematodes and platyhelminthes. There was also a positive correlation between the presence of nematodes and platyhelminthes across all sites - a sequentially Bonferroni-corrected Spearman correlation value of $\rho = 0.0025$ was returned with a significant p-value of $P < 0.05$. No other significant phyla richness correlations were discovered.

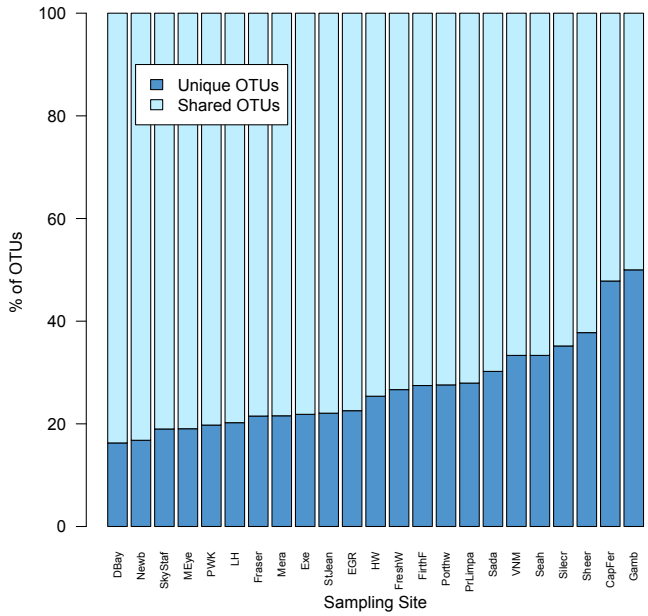

Figure 2.7: Unique and shared OTUs (96%) at each sampling site. A shared OTU is defined as an OTU that appears at more than one sampling site.

In all but one site, Nematoda was the most dominant phylum in terms of number of OTUs. A general phyla richness ranking of Nematoda followed by Platyhelminthes and then Arthropoda is observable with the other, less abundant, phyla more variable in rank (Figures 2.9 and 2.10).

The only association between OTU richness and the environmental variables was between mollusc richness and latitude ($\rho = -0.658$; $P = 0.0006$).

The Mantel-based tests showed that there were significant relationships ($P < 0.05$) between phylum community composition and most variables analysed (the finer grain size - D0.1, seawater surface temperature, geographical distance and latitude) for most meiofauna phyla. Seawater salinity was only significant for Annelida and Tardigrada, whilst the coarser grain size was only significant for the more abundant meiofauna (D0.5 was significant for Platyhelminthes and Nematoda; D0.9 was significant only for Nematoda). None of the variables showed significant associations with any of the protist groups - Rhizaria, Alveolata and Stramenopiles. This information can be found in Table 2.2.

Table 2.3 shows the results of variance partitioning analysis on the factors, latitude, seawater surface temperature, sediment grain size and seawater salinity. In most phyla, latitude and

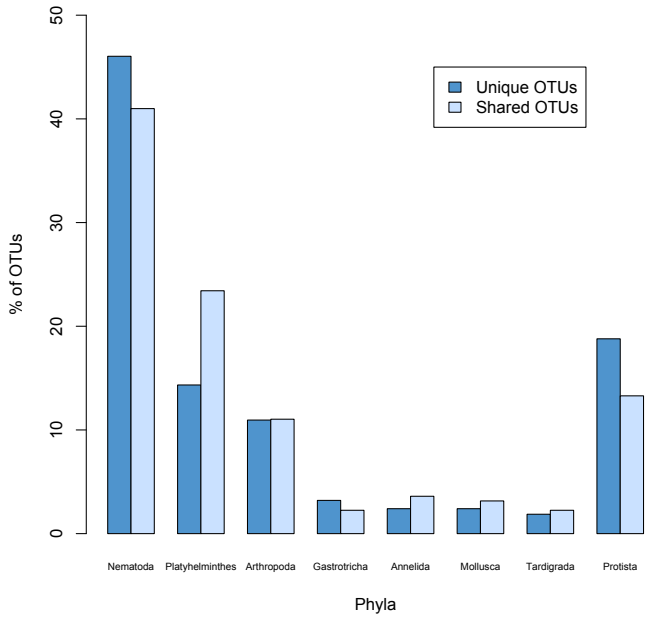

Figure 2.8: 99% OTUs from each phyla expressed as their proportional contribution to the composition of each category (unique or shared). A shared OTU is defined as an OTU that appears at more than one sampling site.

Phylum	D0.1		D0.5		D0.9		SST (°C)		Salinity (%)		Distance (km)		Latitude	
	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P	ρ	P
Nematoda	0.394	0.001*	0.302	0.002*	0.164	0.009*	0.41	0.002*	-0.101	0.778	0.279	0.005*	0.413	0.002*
Platyhelminthes	0.345	0.003*	0.281	0.007*	0.180	0.07	0.380	0.002*	-0.106	0.814	0.320	0.002*	0.416	0.004*
Copepoda	0.289	0.008*	0.195	0.420	0.125	0.12	0.247	0.014*	-0.041	0.634	0.098	0.14	0.168	0.053
Mollusca	-0.067	0.786	-0.039	0.676	-0.039	0.653	0.053	0.282	-0.002	0.492	0.245	0.003*	0.144	0.053
Annelida	0.126	0.125	0.118	0.134	0.057	0.289	0.348	0.005*	0.107	0.018*	0.140	0.083	0.263	0.015*
Tardigrada	0.083	0.184	0.026	0.406	0.003	0.448	0.141	0.072	0.174	0.038*	0.162	0.039*	0.193	0.025*
Rhizaria	0.027	0.380	0.054	0.267	0.002	0.486	0.037	0.341	0.002	0.504	0.002	0.466	0.032	0.314
Alveolata	0.034	0.339	-0.045	0.685	0.074	0.771	0.054	0.240	0.002	0.480	0.072	0.188	0.103	0.132
Stramenopiles	0.023	0.353	0.049	0.269	0.051	0.271	0.012	0.396	0.013	0.508	0.013	0.402	0.031	0.350

Table 2.2: Spearman's correlation (ρ) and Mantel test p-value (P) between community similarity and various environmental variables - grain size (D0.1, D0.5 and D0.9), surface seawater temperature (SST), seawater salinity, geographical distance and latitude - for the main meiofauna and protist (Rhizaria, Alveolata and Stramenopiles) phyla. Significant p-values are marked with an asterisk.

seawater surface temperature account for most of the variance (R^2) in the communities and show very significant ($P < 0.01$ or $P < 0.001$) associations with community structure. The exceptions to this are Annelida, which still returned a significant result for the relationship ($P < 0.05$), and Tardigrada which did not return a significant result.

Conversely, the sediment grain size and the seawater salinity did not explain much of the variance in the communities or show significant associations with community structure in any phyla apart from Gastrotricha. For Gastrotricha, seawater salinity made up almost as much variance as the latitude and seawater surface temperature and showed a significant ($P < 0.05$) association with community structure.

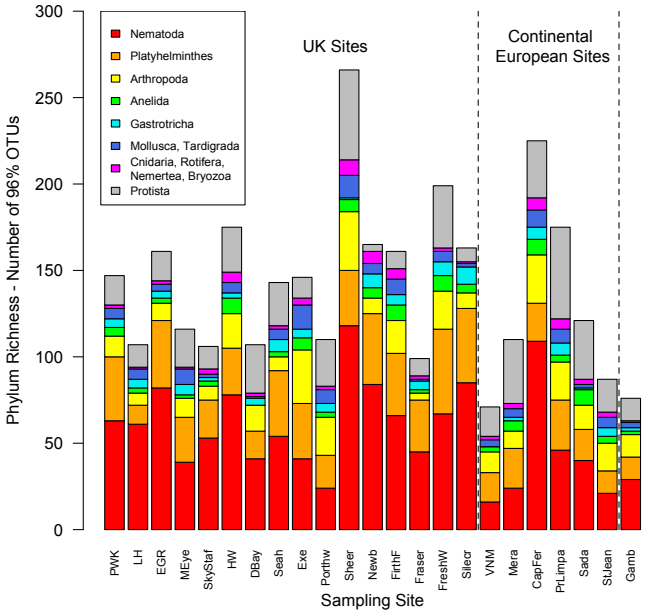

Figure 2.9: Phylum richness at each sampling site, calculated using the number of 96% OTUs.

Phylum	Latitude/SST	S01	Salinity	Residual
Nematoda	0.173***	0.051	0.051	0.725
Platyhelminthes	0.103***	0.058	0.048	0.791
Copepoda	0.127**	0.061	0.039	0.772
Gastrotricha	0.130**	0.046	0.127*	0.699
Annelida	0.099*	0.074	0.055	0.772
Mollusca	0.096**	0.069	0.028	0.806
Tardigrada	0.061	0.054	0.086	0.798

$***P < 0.001$; $**P < 0.01$; $*P < 0.05$; $.\ P < 0.10$.

Table 2.3: Variance partitioning analysis output to show environmental variables and their ability to explain community structure. The R^2 values shown represent the variance attributable to each factor - note that a lot of residual (unexplained) variance is present. The factors S01 and SST are grain size and seawater surface temperature respectively.

Note from Table 2.3 that the effect of latitude and seawater surface temperature were not evaluated at the same time. This is because these two factors are highly correlated with each other and, therefore, gave the same results for R^2 values and significance level. Note also that the residual R^2 values are all around 0.7 to 0.8. This means that much of the variance observable in community structure is unexplained by the environmental and geographical factors that were examined.

The clustering analysis, illustrated in Figure 2.11, indicated that most of the samples taken from the same sampling site were more closely related to each other than they were to samples taken from different sites. An exception to this is one of the samples from Praia Limpa

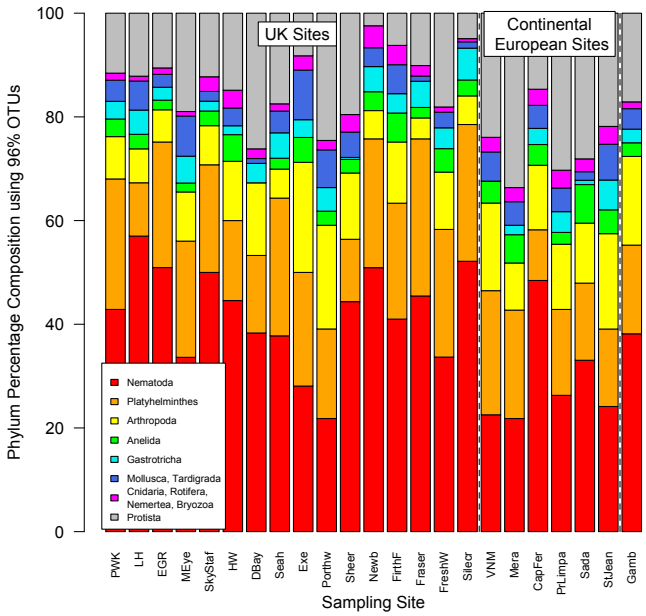

Figure 2.10: Percentage of total OTUs made up from each phylum at each sampling site. 96% OTUs were used.

(PrLimpa1) which bore more resemblance to some UK samples than it did to the other two Praia Limpa samples.

The clusters were tended to be grouped with observable geographical trends. Geographically closer sites were generally more similar than distant sites. Samples from Gambia, as expected, formed an isolated dissimilar cluster. UK sites tended to be more closely related to each other than mainland European sites and vice versa. There were some exceptions, however. For example, the two French sites of St. Jean and Cap Ferret appeared dissimilar to each other, indicating that other factors influence sample composition in addition to geographical distance.

The phylum-specific rarefaction curves shown in Figure 2.12 suggest, due to their relatively steep gradients, that these phyla (Nematoda, Platyhelminthes, Arthropoda, Annelida and Gastrotricha) were under-sampled and that much of their diversity remains hidden.

The data that were found after fitting hierarchical Dirichlet processes to community data in Table 2.4 show that when the community is viewed as a whole, a neutral model is not a good fit. However, if the model is fitted to individual phyla then a neutral model does appear to be an appropriate fit for the majority, especially when applied on a localised scale. More abun-

dant phyla, such as nematodes and Platyhelminthes, are the exception to this and ANOVA results show that there is significant evidence that neutrality is related to the phyla richness on a local scale (Table 2.5). There is also some evidence that neutrality is related to richness across all sites and that it is related to the type of phyla on a local level, with protist groups more likely to follow a neutral model.

Phyla	Classification	96% OTUs	Pseudo p-value (all sites)	Pseudo p-value (localised)
All Phyla	-	1290	0.000	0.000
Alveolata	Protist	52	0.635	0.740
Annelida	Meiofauna	35	0.770	0.805
Arthropoda	Meiofauna	100	0.039	0.337
Cercozoa	Protist	25	0.953	0.902
Gastrotricha	Meiofauna	30	0.474	0.584
Mollusca	Meiofauna	31	0.565	0.544
Nematoda	Meiofauna	413	0.000	0.000
Platyhelminthes	Meiofauna	181	0.014	0.173
Stramenopiles	Protist	100	0.000	0.403
Tardigrada	Meiofauna	20	0.563	0.776

Table 2.4: Pseudo p-values calculated from fitting neutral models as hierarchical Dirichlet processes to community data for different phyla. A pseudo p-value is calculated as the proportion of the likelihood of the fitted metacommunity that exceeded the observed value.

Scope of Neutral Model	Richness (96% OTUs)	Type of Phyla (Meiofauna or Protist)	Residual R^2
All Sites	0.057 .	0.404	0.624
Localised	0.004 **	0.092 .	0.189

$***P < 0.001$; $**P < 0.01$; $*P < 0.05$; . $P < 0.10$.

Table 2.5: ANOVA output to show significance of explanatory variables for the appropriateness of a neutral model.

2.2.4 Discussion

In microbial macroecology it is accepted that, generally, more abundant species are likely to be well dispersed and have high ubiquity levels and, in contrast, rarer species are more likely to be localised (61). The study described in this section shows similar effects on meiofauna with the most abundant phyla (Nematoda and Platyhelminthes) containing numbers of shared OTUs that were disproportionately high when compared with less abundant phyla. This corroborates the previous beliefs about how species' ecology affects dispersal and suggests that more abundant species are more likely to be highly dispersed.

In most samples, Nematoda, followed by Platyhelminthes and Arthropoda were the most dominant phyla. The numbers of the less dominant phyla were more variable from sample to sample with no obvious hierarchy below the aforementioned three phyla. This suggests

that, for more abundant taxa, a neutral model of ecology is not appropriate for the marine benthos, although it may be applicable to the less abundant organisms in isolation. In addition to this, the observed correlation in abundance between Nematoda and Platyhelminthes suggests that they may be competing for the same resources which would promote the idea of an ecological niche.

The most influential factors on community composition were the latitude and seawater surface temperature which are highly correlated with each other because, of course, seawater is warmer closer to the equator. It is apparent that, as in larger organisms, certain meiofauna species thrive in the warmth whereas others prefer cooler temperatures. This characteristic is not noticeable in protist groups (Table 2.2), suggesting that it may not be present in smaller eukaryotic organisms. This is further evidence that meiofauna distribution is niche-driven and that protist groups are perhaps, more affected by spatially limited dispersal..

The above observations are reinforced by the results gained from fitting neutral models as hierarchical Dirichlet processes. These results agreed that neutral models were generally a poorer fit when applied to more abundant phyla and that neutrality was more likely to occur in protist groups, especially when viewed at a local level.

An important question that arises from this analysis is: how would seasonality affect the gathered data? All of the samples were collected during the summer in the Northern Hemisphere but it would be enlightening to see how the communities changed during the year, indeed there have been a number of studies into the effects of seasonality on marine eukaryotic communities (62) (63) (64) all of which show some seasonal change in community structure. This leads to further questions regarding the relative importance of location and climate - how closely would the community of a northerly site in summertime resemble that of a more southerly site later in the year when the temperature has dropped to a similar level?

The type of sediment (with fine silt having a particularly marked effect) has been shown to have an effect on community composition, with that of the phyla Nematoda and Platyhelminthes most influenced by this factor. Continued study in this area may be able to determine requirements and preferences towards different sediment types for each phyla.

The levels of variation between samples taken from the same site were generally very low when compared to variation between different sites. However, there were still some similarities between geographically distant samples that may reflect co-existence between certain species of meiofauna.

As has been shown to be the case in all domains of life (65) (66), levels of similarity in meiofauna communities decreases as geographical distance increases although this effect is not as pronounced in the protist groups. The samples taken from Gambia, the most geographically distant site, showed high levels of beta diversity (diversity between samples) but had the lowest overall community richness. This may be explained by low evenness levels at this site. Other sites showed high levels of both alpha and beta diversity (Cap Ferret, Sheerness and Harwich) which indicates a high rate of turnover at these sites. This could be attributable to unrecorded changes in environmental conditions.

A degree of cosmopolitanism in some species of meiofauna is evident from this study with a proportion of OTUs being shared between multiple sites. Around 40% of OTUs, however, were unique to a particular site which gives evidence for diverse localised communities with a high level of beta diversity. This pattern was noticeable for all phyla and all sampling sites and is similar to that of a previous study (67) which showed 30% of protist taxa as endemic.

The rarefaction analysis indicates that a large amount of species diversity remains undiscovered, showing that the marine benthos is a very diverse and enigmatic environment - due to the evidence of under-sampling inferred from the rarefaction data there is reason to believe that the diversity of the meiobenthos is currently underestimated.

Because of the apparent under-sampling, it is difficult to ascertain whether certain species are genuinely absent from particular sites or if they were merely not sampled. What should be apparent is that there are many low abundance species which have not been analysed adequately and, therefore, their ecology remains more mysterious compared to the better understood ecology of the more abundant species. The limitations on sampling depth mean that this is an unavoidable consequence of this study and others like it.

The richness estimates, using the Chao1 estimator, suggest that there are 2500 meiofauna OTUs in the combined sampled area. Around the UK there was an average of approximately 60 unique OTUs per site with a minimum distance between sites of 20km. Extrapolating this to the 356,000km of the worlds coastline returns an estimate of approximately one million unidentified coastal meiofauna species globally. This estimate is reached using a conservative cut-off of 96% OTUs and is restricted to coastal meiofauna which suggests that the prediction of 2.21 eukaryotic marine species (68) is a major underestimate.

Of the 2500 estimated meiofauna OTUs across all sampling sites, over 830 of these are nematodes. Other estimates for marine nematode richness have predicted that there are 450 species around the British Isles and 1837 species around Northern Europe (47) (69). There

are also reports that 30–40% of free-living Nematoda identified in field surveys of the seas of Europe are new to science (70). Although there is controversy regarding marine species richness, it seems certain that currently richness is underestimated and much of the undiscovered richness is likely to be made up of microorganisms in less explored habitats such as the deep sea and soil (68).

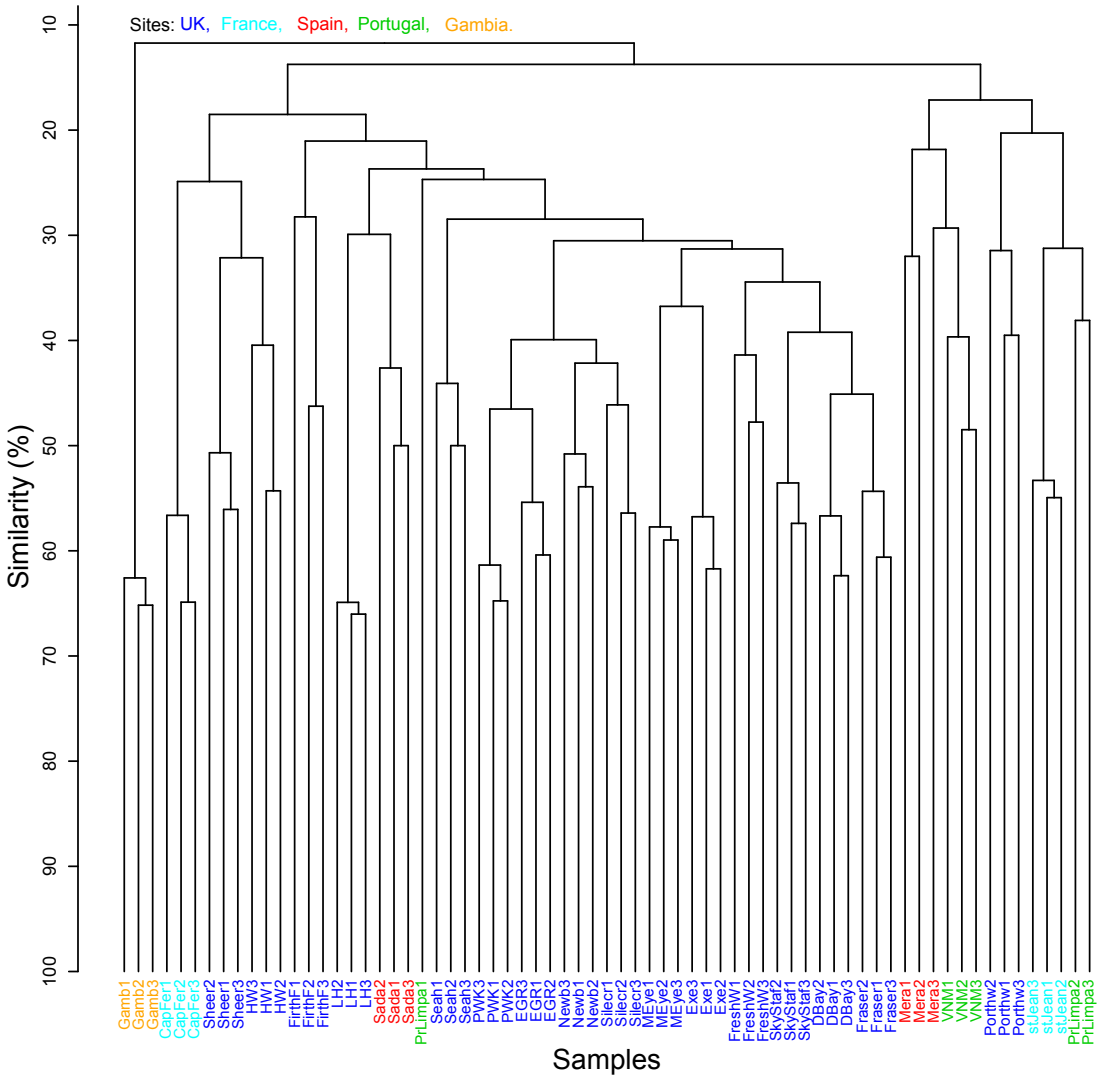

Figure 2.11: Clustering dendrogram to show the similarity of all 69 samples based on Sørensen's coefficient applied to presence/absence data for each sample. The *hclust* function in **R** was used to generate the dendrogram.

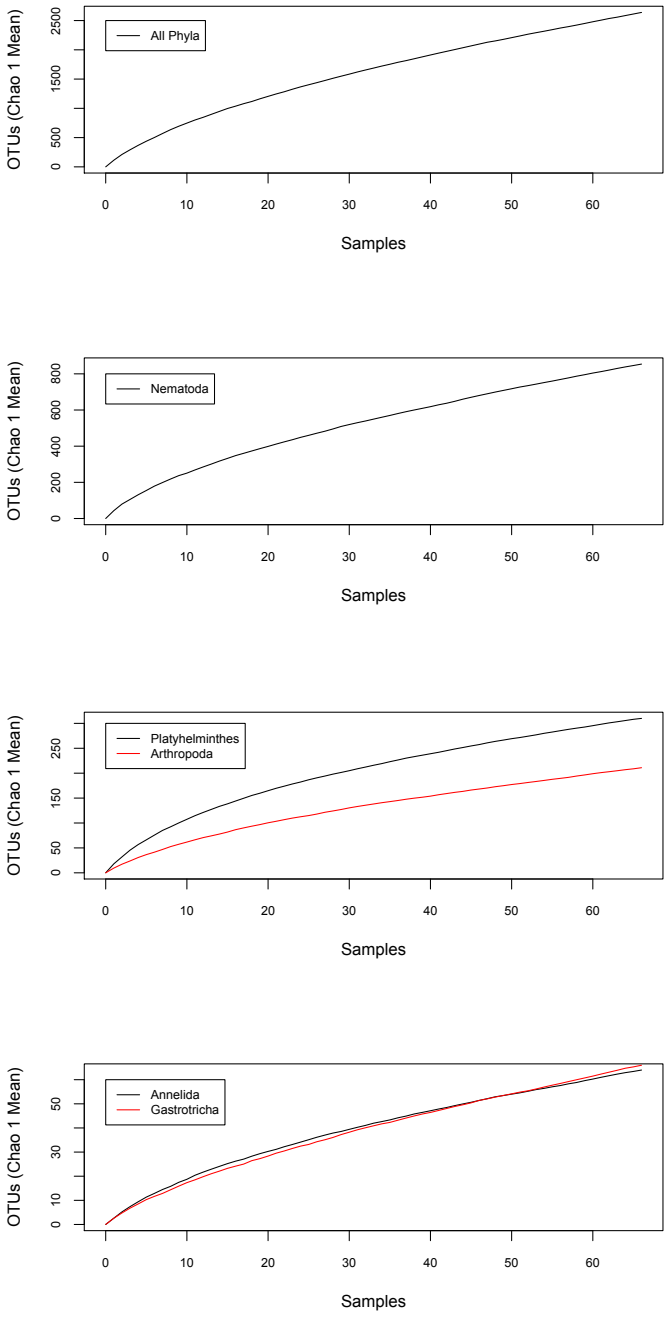

Figure 2.12: Phylum-specific rarefaction curves to show the mean expected number of 96% OTUs (using the *Chao1* richness estimator) against sample size. Curves were estimated from 100 randomisations without replacement using the *specaccum* function in the *Vegan* package in **R**.

2.3 Experiment 2: Investigating the Effects of Genetic Diversity and Sample Richness on Chimera Formation

2.3.1 Introduction

The arrival of next generation pyrosequencing has allowed great progress to be made into the analysis and understanding of prokaryotic and eukaryotic microbial communities (71) (72) (49) (73). One of the barriers to this is the formation of chimeras during PCR (Section 1.4.2) - this is a major issue that affects the reliability of NGS and jeopardises the validity of any conclusions drawn from studies using such technologies. A series of experiments were carried out (46) in which pooled samples of multiple nematodes were sequenced (a variable region within the 18S nSSU gene was chosen) and analysed in order to investigate the formation of chimeras. In addition to this, similar experiments were carried out on 74 samples, each containing a single species of nematode. The exact sequences for these 74 single nematodes were known because they had been found separately using Sanger sequencing (Section 1.2.2).

The proportion of chimeras in datasets generated from nSSU sequencing has been shown to vary from 30–70% (74) (75) (76) and five factors have been shown to influence recombination during PCR (77) (76) (75). These are the number of PCR cycles, PCR extension time, DNA template concentration, *Taq* DNA polymerases and amplicon size. Chimera formation can be inhibited by attempting to optimise the PCR protocol but no method has managed to be sufficiently successful, meaning that post-sequencing chimera detection is the only method available to combat this problem.

The effects of the phylogenetic diversity and richness of a sample on chimera formation have, until now, undergone little investigation barring a small study which was carried out on the effects of sequence similarity on chimeras using cloned 16S rRNA genes and mixed bacteria genomic DNA (76) (78). This study did not consider sample richness and pre-dated the current second-generation sequencing perspective of amplicon pool diversity.

The main goals of the research presented in this section were to, firstly, investigate the effect of sample richness, evenness and genetic diversity on the formation of chimeras and to link this to diversity estimates. The second goal was to investigate the role of variation within the amplicon sequences, and also the variation of the secondary structure of the nSSU molecule on chimera formation.

2.3.2 Materials and Methods

Sample Preparation

The sequences of 74 Sanger-sequenced individual nematode species were blast aligned to a contemporary Nematoda phylogenetic framework (79). In order to divide these sequences into pools of closely related species and distantly related species, an alignment was created using ClustalX and the pairwise distances (p-distance) between sequences were calculated using MEGA-4.1 (80). Closely related pools were formed from sequences with mean p-distance (MPD) of less than 25% - referred to as 'phylogenetically close'. Distantly related pools were formed from sequences with MPD of greater than 40% - referred to as 'phylogenetically distant'. In all, 30 pools were formed - 15 phylogenetically close pools (5 with 12 species, 5 with 24 species and 5 with 48 species) and 15 phylogenetically distant pools of the same makeup.

DNA Extraction and Preparation

DNA was extracted from DESS-preserved nematodes (56) using a DNeasy blood and tissue kit (Qiagen Inc). The DNA was eluted in 40µl of AE buffer and stored at -20°C. A Nanodrop spectrophotometer was used to quantify DNA extracts from all individual nematodes which were then diluted to 0.5ng/µl.

PCR Amplification and Sequencing Analysis

The forward primer:

SSU_FO4 (5'-GCTTGTCTCAAAGATTAAGCC-3')

and the reverse primer:

SSU_R22 (5'-GCCTGCTGCCTTCCTTGGA-3')

were again used to amplify approximately 450bp of the V1-V2 regions of the nuclear small subunit rDNA (18S rDNA).

Fusion primers were developed (49) and PCR amplification reactions and the thermocycle for the targeted region were optimised using 0.25ng/µl of genomic DNA template in three 40µl reactions, where *Pfu* DNA polymerase (promega) was used for each of the phylogenetically close and distant nematode pools and all individual DNA extracts.

PCR thermocycling was initiated with a 2 minute denaturation step at 95°C which was followed by 35 cycles - intended to optimise the number of chimeras formed (74) (76) (78) - of 1 minute at 95°C, 45 seconds at 55°C and 3 minutes at 72°C for each cycle and a final extension of 10 minutes at 72°C. Negative controls using pure water only were applied for all amplification reactions.

Top Vision™ LM GQ Agarose (Fermentas) on a 2% gel was used to undertake the electrophoresis of the triplicate PCR products and the QIAquick Gel Extraction Kit (Qiagen) was used to purify the expected 450bp fragment in accordance with the manufacturer's instructions. An Agrilent Bioanalyser 2100 was used to quantify all purified PCR products before they were all diluted to the same 10ng/μl concentration.

Sequencing was performed at Liverpool University's Centre for Genomic Research, UK, using a 454 Roche GSFLX. All PCR amplifications were sequenced in a single direction (A-Amplicon) with the single nematodes sequenced on a quarter of a plate and the pooled nematodes sequenced on three quarters of a plate.

Denoised Reads and Chimera Detection

AmpliconNoise was used to remove the noise from the resulting amplicons following the filtering, flowgram and clustering steps described in Chapter 1 and Perseus was used, also as described, to identify the chimeras.

The output from Perseus gives the most likely break point for each chimera based on its two identified parent sequences - calculated by minimising the number of differences from each contributing parent when both are aligned with the chimera. These break points were standardised for the whole dataset by forming a four-way alignment of each chimera, its two parents and a reference sequence (*Caenorhabditis elegans*) using ClustalX (16). The position of each break point on the reference sequence was recorded to give a standardised break point. The frequency of each standardised break point could then be recorded to assess which regions of a sequence were most susceptible to chimera formation.

The potential role of the 18S rDNA amplicon region's secondary structure on chimera formation was investigated using MFold RNA-folding software (81).

Generation of OTUS

OTUs were generated with a 99% identity cut-off using a complete linkage clustering algorithm as described in Section 2.2.2. The number of OTUs in each sample was used as an estimate of taxa richness and the effects of this metric on chimera formation was investigated.

Species Diversity and Nucleotide Diversity

As has been discussed in Chapter 1, species diversity or, more accurately for the analysis in this chapter, OTU diversity is measured using the Shannon index,

$$H' = -\sum_{i=1}^{S} \{p_i \ln(p_i)\}$$

where S is the total number of OTUs and p_i is the probability of a randomly chosen individual belonging to OTU i.

Nucleotide diversity was also calculated using this index. An alignment of all good (non-chimeric) sequences with the *C. elegans* reference sequence was formed and a value of H' was found for each position on this alignment. In this case, $S = 4$ represents the number of possible nucleotides and p_i is the proportion of nucleotide i at the position in question.

Analysis of Variance

To analyse the relationship between explanatory variables (e.g. relatedness, number of individuals in experiment, number of reads, diversity of sample) and the overall chimera percentage, analysis of variance (ANOVA) was used.

A linear model was fitted to the data, giving a multiplicative coefficient for each explanatory variable. ANOVA was carried out to determine a p-value for each explanatory variable, i.e. the probability that its coefficient is equal to zero. In other words, a small p-value is evidence that the associated explanatory variable has an effect on the chimera percentage. Unnecessary variables were removed and the model was refitted to give an accurate ANOVA table.

Three models that were analysed were:

1. Chimera percentage versus relatedness, number of individuals and number of reads.

2. Chimera percentage versus relatedness, number of OTUs and number of reads.

3. Chimera percentage versus relatedness, Species Diversity and number of reads.

These models had to be analysed separately because there is obvious dependence between number of individuals, number of OTUs and Shannon index. The linear modelling (*lm*) and ANOVA functions found in **R** were used for this analysis.

2.3.3 Results

The ANOVA output suggested that Relatedness (p-value 3.5×10^{-6}), Species Diversity (p-value 4.5×10^{-4}) and the number of OTUs (p-value 7.9×10^{-3}) all had a significant effect on the number of chimeras formed.

Output from Perseus indicated that the amount of chimeras present ranged from around 14% to 60% of the total sequences - see Table 2.6. From the ANOVA results, it can be seen that the two main causes of this variation are species diversity and species relatedness. This is illustrated in Figures 2.13 and 2.14 where it can be seen that the more distantly related pools produced more chimeric sequences (Figure 2.13) and also more chimeric reads as a percentage of the total number of reads (Figure 2.14). A clear positive correlation between species diversity and the chimera percentage can also be seen in both cases. This correlation was also observed by (82) and (78) using bacterial data.

An important influence on the formation of chimeras is the position where PCR fails, thus forming the fragment of DNA from which the chimera is generated. Clearly, if a sequence contains a region that is more susceptible to PCR failure then the probability of chimera formation is increased. Figure 2.15 shows the relationship between nucleotide diversity at a given position and the break point frequency at that position. There is a negative correlation between the two variables, that is, more conserved regions of the sequence tend to result in break points and regions were nucleotide diversity is higher contain fewer break points.. The relationship was shown to be significant, with a sufficiently small probability (p-value 3.9×10^{-4}) of there being no correlation.

The majority of break points occurred between positions 80 and 200 on the alignment. Figure 2.16 shows the break point frequency (bottom graph) and nucleotide diversity (top graph) at each of these positions. The relationship can be seen in the way the peaks in the top graph line up with the troughs in the second graph and vice versa.

Break point histograms divided into the six different pools (closely and distantly related

pools of 12, 24 and 48 nematode species) are shown in Figures 2.17 and 2.18. These demonstrate similar break point distributions for all pools, particularly the small peak around position 100 and a larger peak around position 170 on the alignment with the reference sequence.

Related	#Species	#OTUs at 99%	#Sequences	Chimera%	#Reads
Close	48	87.6	138.40	35.60	13882.20
Close	24	40.4	63.20	34.55	3809.00
Close	12	35.8	42.80	14.57	6159.80
Distant	48	63.2	161.00	58.98	5657.80
Distant	24	53.6	119.00	53.57	10134.20
Distant	12	34.4	58.20	39.93	7638.20

Table 2.6: Data for each experiment. Values shown are the means of the five repetitions.

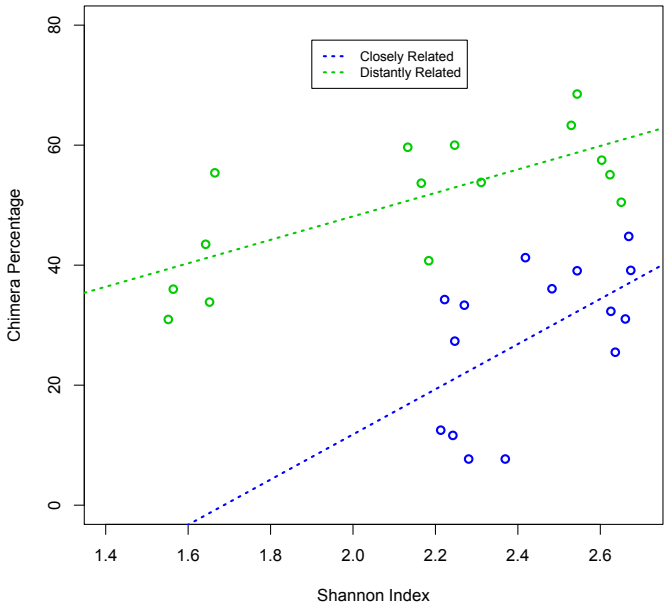

Figure 2.13: Chimera formation against species diversity shown for closely related and distantly related pools. The species diversity for each sample was calculated as the Shannon index of the denoised data.

2.3.4 Discussion

The number of chimeras generated (up to 60% in some datasets) confirm that 35 rounds of PCR do tend to generate a large number of chimeras, as has already been claimed by previous studies (76) (82) (83). This result reinforces the necessity of running chimera detection software in order to process sequencing data to a state where they are fit for analysis. Failure

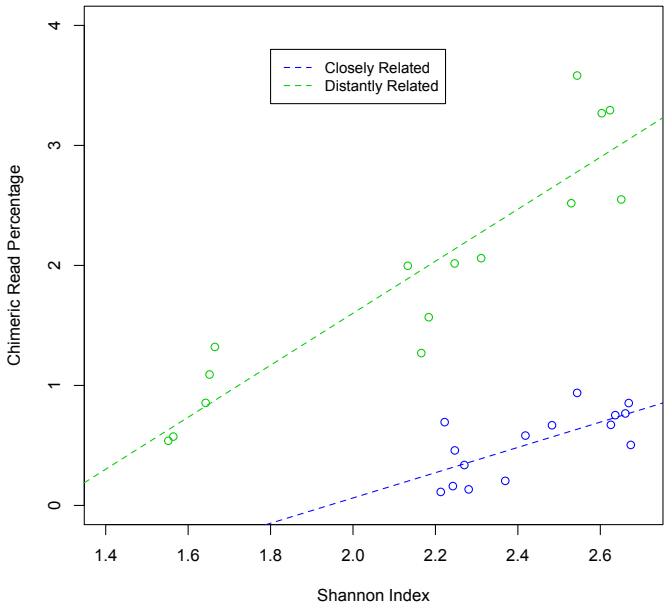

Figure 2.14: Chimeric read percentage against species diversity shown for closely related and distantly related pools. The species diversity for each sample was calculated as the Shannon index of the denoised data.

to do so will undoubtedly lead to over-inflated OTU richness and erroneous diversity estimation in environmental samples (18) (84) (85).

The number of OTUs, after chimera removal, found in each sample were generally around double the number of nematodes chosen for that sample. Reasons for the extra OTUs could be the presence of undetected chimeras or possibly genetic material from other organisms found on or in the chosen nematodes (as prey). Another possibility is the fact that organisms often contain multiple copies of heterogeneous nSSU genes (86).

The impact on the dataset of these multi-copy nSSU genes, all single nematodes were amplified with unique MID-tag sequences. Of these amplifications, 61 were single copy 18S rDNA and 11 were double copy, however all taxa were represented by a similar number of taxa in PCR reactions suggesting that the presence of multi-copy nSSUs had little effect.

The significant results demonstrating that distantly related pools of nematodes and pools of nematodes containing more species tend to yield more chimeras clearly give strong evidence that phylogenetic diversity and species richness are contributing factors to chimera formation in nSSU amplicon pools. Further evidence to support this hypothesis can be seen from the comparison of the Shannon diversity indices of the samples. Samples with higher Shannon indices tend to produce more chimeras.

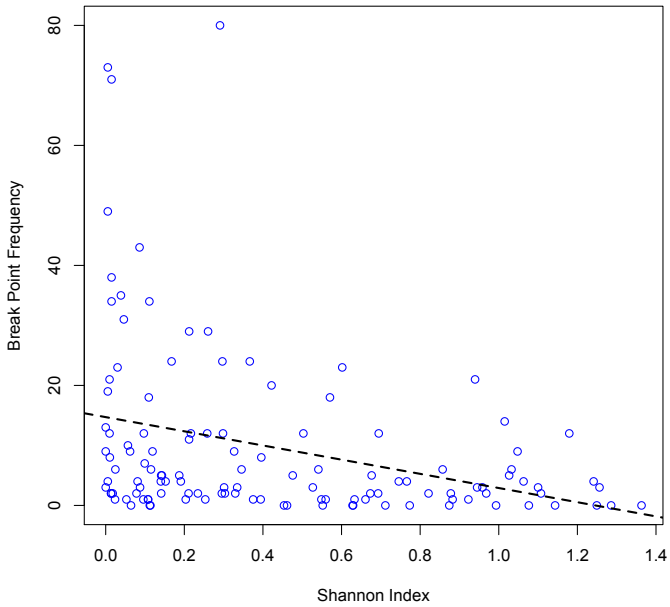

Figure 2.15: Nucleotide diversity (Shannon index) against break point frequency. A four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded. The nucleotide diversity was calculated using the Shannon index at each point on a multiway alignment between all good sequences and the *C. elegans* reference sequence.

The investigation into the effect of nucleotide diversity on chimera break points yielded results which show that regions of lower nucleotide diversity are more likely to instigate chimera synthesis. These results compare favourably with studies on the bacteria 16S rRNA gene which found correlations between sequence similarity and chimera formation to exist (77) (74) (78). A likely explanation for this is that more conserved regions will be better equipped to bind with a PCR fragment acting in lieu of a primer, as they are more likely to share matching sequence segments with the fragment.

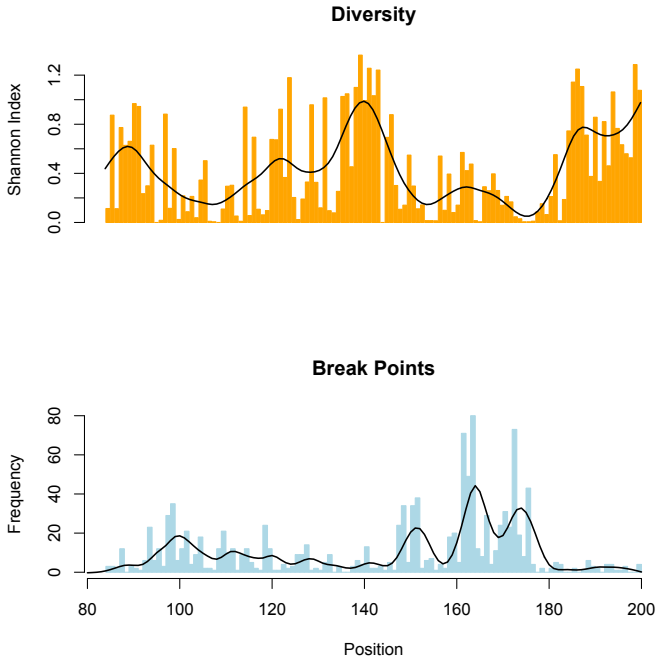

Figure 2.16: Nucleotide diversity (Shannon index) and break point frequency plotted against position of break point. The lines show the same information with smoothed data. A four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded. The nucleotide diversity was calculated using the Shannon index at each point on a multiway alignment between all good sequences and the *C. elegans* reference sequence.

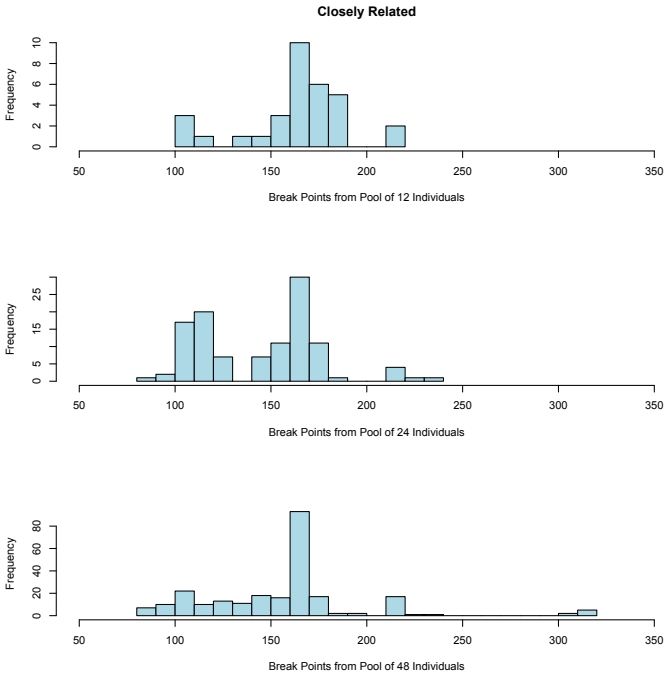

Figure 2.17: Break points for closely related nematode species. A four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded.

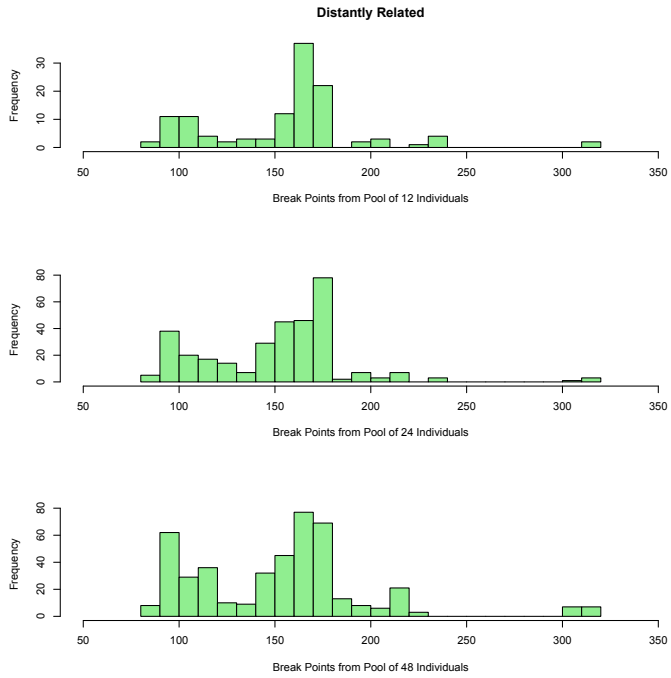

Figure 2.18: Break points for distantly related nematode species. A four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded.

Chapter 3

Modelling the PCR Process to Simulate Realistic Chimera Formation

3.1 Introduction - Why is a New PCR Model Required?

Polymerase Chain Reaction (PCR) is the principal method of amplifying target DNA regions and, as such, is of great importance when performing microbial diversity studies. An unfortunate side effect of PCR is the formation of unwanted byproducts such as chimeras.

The main goal of the work covered in this chapter is the development of an algorithm that simulates realistic chimeras for use in the testing of chimera detection software and for investigations into the accuracy of community structure analyses. Experimental data has offered insights into identifying factors which may cause the formation of chimeras and has provided evidence of how influential these factors can be. This chapter makes use of some of this evidence in order to build a model with which to simulate the PCR process. This model helps to better explain the formation of chimeras and is therefore able to provide aid to future studies that intend to use PCR.

As is discussed in the following section (Section 3.1.1), whilst a number of PCR models exist, there is a sparsity of models built for the purpose of artificial chimera generation. Those that do simulate chimeras, do so in such a way that the amount produced is based on the user's desired number of chimeras. A more realistic model would rely on the composition of the input sequences and values of parameters modelling PCR conditions to drive chimera generation - the number of chimeras produced and their composition should be dependent on the input and not predetermined. Simulation software meeting these requirements would

be very welcome.

An advantage of simulated data is the presence of complete information - because the input data is known then it is possible to separate the output data into chimeras and good reads with 100% accuracy. If, then, the simulation proves to be realistic enough it will be extremely useful for testing chimera detection software without the required time and expense of experimental data.

It has been claimed that the leading chimera detection tools, Perseus and UCHIME, can detect nearly all chimeras in a dataset with few false positives (18) (20) but just how confidently can these assertions be made? Both Perseus and UCHIME were tested on mock community datasets with good results, however, it would be desirable to see how the results would compare if they were tested using a dataset with a more realistic community structure, chimera frequency and chimera composition. The models formulated in this chapter are used in Chapter 4 to generate *in silico* datasets designed for this purpose.

If chimera removal software does not perform as well as has been imagined then this would be cause for concern. The presence of undetected chimeras in datasets could give a false picture of community structure, likely overestimating richness and diversity levels, and would ultimately add a significant degree of uncertainty to the findings of any research that has been carried out on such data.

The findings from Chapter 2 show that chimera formation is a complicated process affected by a number of different factors such as relatedness, species diversity and nucleotide diversity. All of these factors contribute and interact to influence the formation of chimeras in ways that are difficult to understand using experimental data alone. It would, therefore, be very interesting to see whether a model designed to simulate chimera formation could help to explain how this complex system works. If a model could somehow incorporate all of these factors, then the different interactions between them could be explored and it may be possible to determine which factors have the most influence on the formation of chimeras.

There is the possibility that other, as yet unknown, factors could also contribute to the level of chimera formation. In addition to this, the amount of randomness involved is not understood. A good model of the PCR process, designed specifically with chimera formation in mind, would allow comparisons to be drawn between experimental and simulated data. This would allow improvements to be made to chimera identification and noise removal techniques.

In conclusion, there is clearly a need for a PCR model that better simulates chimera gen-

eration.

3.1.1 Existing Models of PCR

Many different studies into the simulation of PCR have been carried out in the past. Differing limitations, areas of study and goals relating to the usage of these simulations have led to varying levels of complexity and various different applications.

Some existing PCR simulators operate by selecting target regions from a set of longer genome sequences when given the primer sequences as input and return the required amplicon sequences as output. Rubin et al. (87) present such a model which is designed to investigate the production of non-targeted PCR products using a simple algorithm that matches primer sequences to suitable template DNA sequences based on a maximum mismatch threshold. The study concludes that, according to the results of the simulation, more unwanted PCR products are formed in practice than predicted by the model.

Another similar PCR simulator is *ecoPCR* (88) which takes a primer pair as command line input and makes use of the Wu-Manber algorithm (89) for pattern searching. This algorithm compares two strings and indicates whether or not the longer string contains a substring that is "approximately equal" to the shorter string. In other words, two strings are treated as identical if they are within a specified *Levenshtein distance* (90) of each of other. The Levenshtein distance is, in basic terms, a measure of the number of insertions, deletions or substitutions required to convert a given string into a target string. In the context of simulating PCR, the Wu-Manber algorithm is used to search for the optimal region of a given sequence with which to bind a primer. Output from ecoPCR includes the amplicon sequence, its length, the number of mismatches on each primer and various taxonomic information relating to the sequences.

There are also several websites which offer PCR simulation via the input of sequences and primers directly into the user's web browser as well as changing variables relating to PCR conditions. Examples of such websites are `cybertory.org` (91), `bioinformatics.org` (92) and `amnh.org` (93). The usage of these tools is generally limited to data containing fewer input sequences.

Primer Prospector (94), whilst not designed specifically as a PCR simulation tool, may be used in the same way as much of the software described in this section. The tool assesses the ability of a primer pair to act on a dataset of sequences and outputs statistics based on the proportion of these sequences that can be expected to amplify as well as a file containing all

of the amplicons generated.

As is the case with those outlined so far in this section, the majority of available tools simulate PCR by extracting the targeted sequence fragments from the reference. They predict probable PCR products and generate statistics about potential mismatch locations and primer efficiency but they do not imitate a PCR process. An exception to this is *Grinder* (95) which produces simulated PCR amplicons with chimeras and single-base PCR errors included. Chimeras may be generated from an input parameter specifying the percentage of chimeras required and, similarly, the number of PCR errors can be controlled by inputting the required mutation rate and distribution. In Grinder, a chimera may be generated in one of two ways - the first method is randomly selecting a pair of parents and a random break point and the second is similar to the method used by CHSIM. Chimeras are then randomly added to the output data based on the required chimera proportion.

CHSIM is the name of the chimera simulation algorithm which was used to generate chimeras for the purpose of testing UCHIME (20). The algorithm selects parent sequences which share an identical sub-sequence (*k-mer*) of given length, this k-mer is used as the crossover section between the two parents (i.e. the break point is contained somewhere within this section). Chimeras are generated at random, weighted in favour of those containing the most abundant k-mers present in the pool of potential parents. This is intended to make break points more likely between similar sequences in regions of high sequence similarity. A preset number of chimeras are generated in this way and added to the original pool of parents after each simulated round of PCR.

3.1.2 Choosing a Good Model

In order to choose a good model for any procedure, several things should be considered such as the model's complexity as well as the parameters and input required for the model. The number of different variable parameters will impact on the model's complexity and it may be decided that it is best to ignore certain variables in order to simplify the model. It is important to correctly identify the sources of variation that affect the process in practice and to model these realistically using appropriate methods. One example of this is the selection of appropriate probability distributions from which to draw random variables.

A good model should also be easy to implement and run quickly enough so as to be practical. The functionality of the model should be expressible in the form of an algorithm that can be implemented in code. When implementing the algorithm, compatibility with existing software and file formats (for input and output) must be taken into consideration. If large

amounts of data are to be processed then it is desirable to use an algorithm that minimises the number of calculations in order to reduce the running time. Sometimes it may be better, or even necessary, to forfeit some accuracy in order to produce a faster algorithm.

Most factors that should be considered when choosing a good model will have an effect on its complexity and often a trade-off between complexity and accuracy will be necessary. A simple model is more desirable if it is as effective as more complicated models. However, if a model is oversimplified then there is a danger that its output will be unrealistic. For example, a very simple model of PCR would be to take as input the initial abundance of each DNA sequence and increase this amount based on the number of PCR rounds, such that

$$a_{new} = a_{old} \times 2^n$$

where a_{old} and a_{new} are, respectively, the original and resultant abundances of the sequence and n is the number of PCR rounds. To calculate the new abundance, the old abundance is multiplied by a factor of two raised to the power of n because each sequence splits into two new sequences during each round of PCR.

Output from this model will not be useful in practice because it does not take into account the randomness and errors inherent in PCR amplification. In particular, it ignores the facts that amplification is not 100% efficient and that the amplification step can fail before completion, creating artefacts that further complicate matters.

3.2 Methods

Chimera break point distributions taken from experimental and simulated data were compared using the *two sample Kolmogorov-Smirnov test* (96). This test returns a p-value to indicate the probability that the two samples are similarly distributed. This means that an insignificant p-value (typically $p > 0.05$) will reveal no information about the similarity of the two sample distributions but it can be concluded that they are similar enough that there is no obvious distinction.

The Kolmogorov-Smirnov test is typically used for samples with continuous data, however it has been adapted for discrete samples in the **R** package, *dgof*, and is therefore applicable for the analysis of break point distributions. Before each Kolmogorov-Smirnov test was carried out, the larger of the two samples being tested was sub-sampled to the same size of the smaller. Because different sample selections give different p-values, the process was

repeated 100 times in each case and the mean p-value was taken.

Correlation between break point frequencies occurring in experimental and simulated output was assessed using *Pearson's correlation coefficient*, r_{XY}, which is calculated using the formula,

$$r_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right) \left(\frac{Y_i - \bar{Y}}{s_Y} \right)$$

where n is the number of observations in each sample (both samples must contain the same number of observations), X_i and Y_i are the break point frequencies at position i in sample X and sample Y respectively, $\bar{X}$ and $\bar{Y}$ are the mean break point frequencies across all positions in sample X and sample Y respectively and s_X and s_Y are the sample standard deviations. As for the Kolmogorov-Smirnov test, the two samples were sub-sampled to the same size 100 times and the mean of the 100 different Pearson's correlation coefficients was recorded.

Similarity in nucleotide composition between simulated chimeras and chimeras generated experimentally was assessed using the 'global search' function in *USEARCH* (97). One dataset of chimeras (e.g. experimental chimeras) was used as a query dataset to be searched against a reference dataset (e.g. simulated chimeras). Sequences in the query dataset were paired with the most similar sequence in the reference dataset and a similarity score was recorded (number of matching nucleotides divided by alignment length).

3.3 The PCR Process

This section presents a summary of PCR as a procedure, the steps of which must be emulated to develop a realistic model of the process. The PCR process is also summarised visually in Figure 3.1.

In order to prepare a sample for sequencing, an amplification step is carried out using Polymerase Chain Reaction (PCR). Thermal cycling is used to repeatedly melt and cool the DNA. When a strand of DNA is copied, this copy can then also be copied; this leads to an exponential amplification effect. PCR is used to amplify a particular target region of the DNA - this is selected using primers (small pieces of DNA, complementary to the target region).

The process typically involves 20-40 cycles of the following steps (2^{40} gives approx 10^{12} copies):

1. Denaturation – this step takes place at temperatures between 94 and 98°C for around

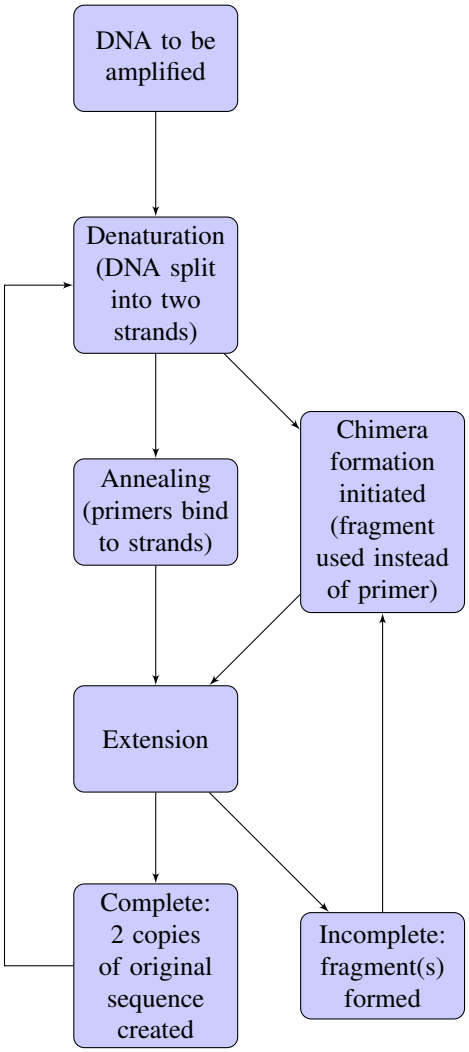

Figure 3.1: The PCR process.

20 to 30 seconds. Hydrogen bonds are broken to split the DNA into two strands.

2. Annealing – the temperature is reduced to 50-65°C. The primers bind to both single strands of DNA. Hydrogen bonds are only able to form when there is a close match, ensuring that the primers are annealed to the correct region.

3. Extension – the temperature is adjusted depending on the polymerase used. Nucleotides are attached to complete the DNA strands. These strands can now be copied in the same way as the original.

Forward and Reverse Primers – After the annealing step, when the DNA molecule has been split into two strands, the primer binding onto one of these strands is called the *forward primer*. Extension can only occur in the 5' $\rightarrow$ 3' direction, this means that the primer binding to the second strand of the complementary pair must induce extension in the opposite direction to the first. A different primer, the *reverse primer*, must be used for this.

Chimera Formation – Chimeras can be formed when the PCR extension step is incomplete. If PCR fails at a certain point then an incomplete sequence of DNA is produced, this fragment can act as a primer for a different sequence in another round of PCR. This has the effect of forming a sequence which is really a combination of two or more different partial sequences. The proportion of chimeras present varies from dataset to dataset. Some datasets can be comprised of 90% chimeric reads. This is obviously a large problem that is addressed using noise removal software.

3.4 Model 1

3.4.1 Model Outline

The repetitive cyclic nature of PCR suggests that an intuitive model is an iterative procedure with the same steps being repeated for every simulated round of PCR. The basic input information that will be required are the number of rounds of PCR, the primers to be used, the DNA sequences to be amplified and their initial abundances.

There are two factors that drive the way PCR progresses. The first of these is the rate of failure of PCR, when the two parts of a DNA strand do not combine with primers or fragments to begin amplification and, instead, simply recombine with each other. This failure rate will depend on the relative concentrations of sequences, primers and fragments and can be calculated each round. The second factor is the rate of failure during extension. Parameters used in this model should be chosen with these factors in mind.

- Set initial pool of sequences and abundances.
- Set empty pools of forward and reverse fragments.
- Set primer sequences and initial primer abundances, a_p.
- Set λ.
- Set number of PCR rounds.
- For each round of PCR, r:
  - Recalculate α from sequence, fragment and primer abundances.
  - For each sequence s (abundance a_s and length l_s):
    - For each fragment/primer f (abundance a_f):
      - Calculate PCR failures as a Binomial(a_s,α) random variable.
      - Decrease a_s by this amount.
      - Record differences and break point for fragment f.
      - Calculate weight, W_f, for fragment f.
    - Generate vector of quantities $[c_p, c_1, \dots]$ as Wallenius$(X_s, [a_p, a_1, \dots], [W_p, W_1, \dots])$ random variable.
    - Add chimeras to pool of sequences.
    - Repeat for reverse fragments/primers.
    - For each fragment/primer, f (length l_f) and for each potential break point, $b = (l_f + 1) \dots (l_s - 1)$:
      - Generate number of fragments of length b as a Binomial(c_f, λ) random variable.
      - Add these fragments to fragment pool.
      - Decrease c_f.
      - Double the remaining value of c_f - successful amplification.
    - Repeat for reverse fragments/primers.
- End of algorithm.

Figure 3.2: Simera algorithm for Model 1. λ is the rate of failure during PCR extension at each nucleotide point on a sequence. α is the PCR failure rate and is calculated using the formula in Section 3.4.1. The fragment weightings, W_f, are calculated using the formula in Section 3.4.1.

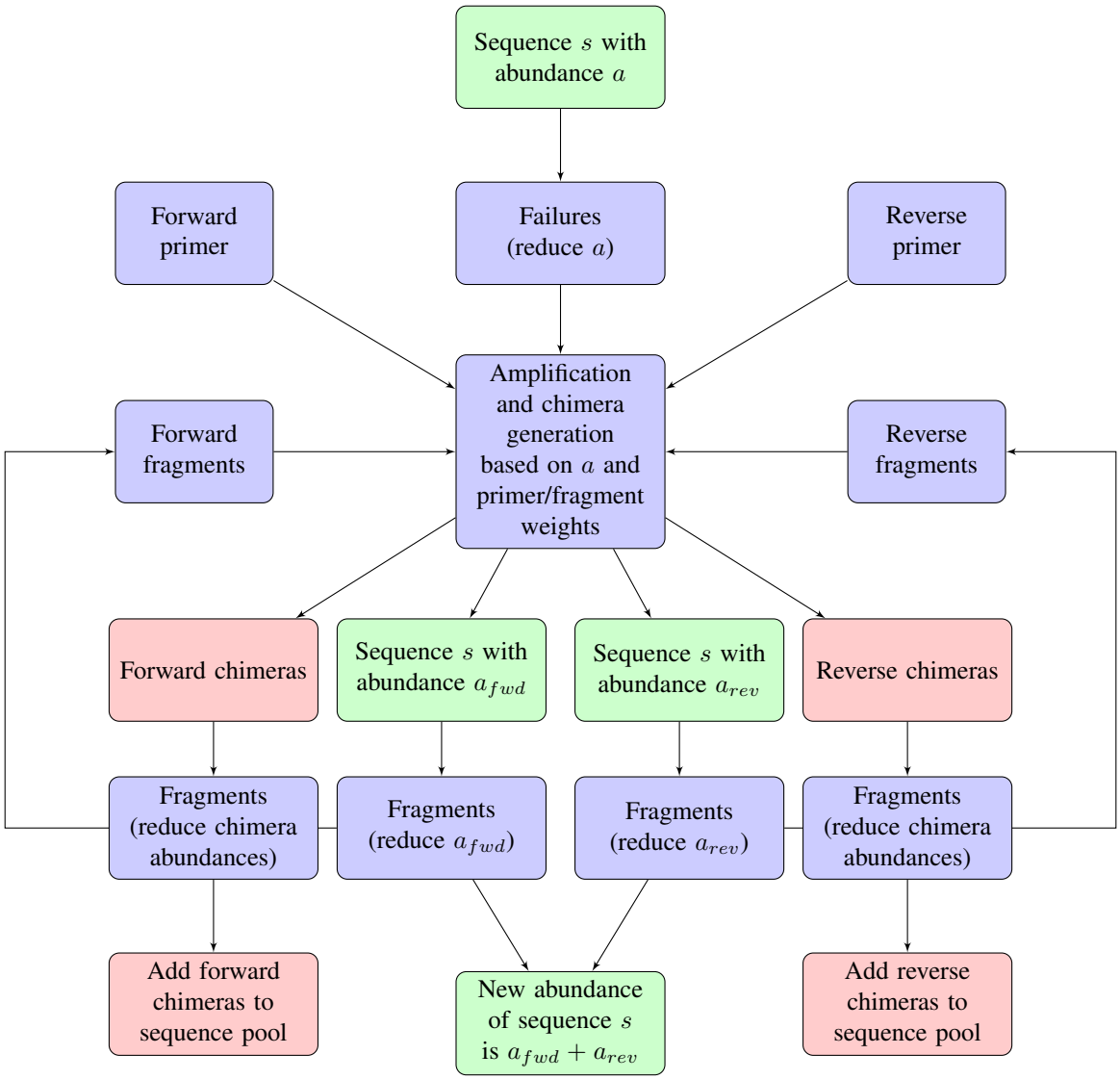

Figure 3.3: PCR simulation using Model 1. This process is performed for all sequences, s, and repeated for the desired number of PCR rounds.

One possible approach for modelling PCR, and the approach used in this chapter, is to use integer values for the abundance of each sequence. This means that a sequence will be treated in the model as an individual strand of DNA and allows the model to be closely analogous to the actual process. Because of this, discrete probability distributions, such as the binomial distribution, will be required to generate the random variables necessary for the model.

The steps which make up the algorithm for the model are described in detail in the remainder of this section. The complete algorithm is referred to as *Simera*, a portmanteau of the words "simulation" and "chimera". The Simera algorithm is presented in Figure 3.2 and visualised in Figure 3.3.

Assumptions

For this model it is assumed that failures during the PCR extension step will occur at a fixed rate. That is, extension is equally likely to fail regardless of the position on the DNA strand being amplified and regardless of the nucleotide content at this position.

Complete PCR failures - PCR failures without any extension - are assumed to be dependent on the relative primer abundance which will decrease in later rounds.

It is assumed that the ability of a fragment to act in place of a primer is directly affected by its degree of similarity to the true primer. Further to this, it is assumed that fragments generated by forward extension ($5' \rightarrow 3'$) may only act in place of forward primers and those generated by reverse extension ($5' \leftarrow 3'$) may only act in place of reverse primers.

Input Parameters

1. n – The number of rounds of PCR to be simulated.
2. λ – This parameter is the rate of failure, during the extension step, at each nucleotide on a sequence. It is used to determine if the first nucleotide is duplicated, then the second, etc. until the entire sequence has been amplified. If amplification fails at any point then a fragment is produced. λ is a probability between zero and one, and should typically be very small. λ may depend on PCR conditions so should be variable from dataset to dataset.

Sequences

A list of initial sequences and their relative abundances shown as integer values are required as input for the Simera algorithm. The sequences will each be a string of DNA nucleotide codes [A,C,G,T] and only the region selected for amplification need be included. In practice, for implementations of the model, a fasta file is a good way to represent these data.

Primers

Information about the forward and reverse primers must be also be supplied as input. The primers will be a string of DNA IUPAC codes [A,C,G,T] and will typically be about 20 base pairs long. Unlike the DNA sequences, primers may also contain ambiguous IUPAC codes [R,Y,S,W,K,M,B,D,H,V,N] which each represent two or more of the four specific DNA nucleotides. For example, a primer containing the code M in the first position actually represents a collection of primers where 50% contain the A nucleotide and 50% contain the C nucleotide in the first position.

These codes are included in primers because they are more versatile and can, therefore, be better at selecting sequences which have a high degree of nucleotide variation at certain points. The ambiguous IUPAC codes and the nucleotides which they represent are shown in Table 3.1.

As input data, the abundance of each primer is also required. This should be an integer value and should be greater than the number of primers required to perfectly amplify all sequences for the given number of rounds, n. Therefore, if the initial sequence abundance is a_{seq} then the initial primer abundance should be

$$a_{primer} > a_{seq} \times 2^n.$$

Fragments

Two lists of sequence fragments are also required. Initially these are empty but, during the simulation, fragments will be generated and recorded. The first pool is a pool of forward fragments - those generated from forward primers - and the second is a pool of reverse fragments. The abundance of each fragment is defined the same way as in the pool of sequences. During the simulation the primer abundance and the abundance of incomplete sequence fragments will be used to calculate the probability of chimera formation where a fragment is selected in place of the primer.

Code	Proportion of A	Proportion of C	Proportion of G	Proportion of T
R	$1/2$	0	$1/2$	0
Y	0	$1/2$	0	$1/2$
S	0	$1/2$	$1/2$	0
W	$1/2$	0	0	$1/2$
K	0	0	$1/2$	$1/2$
M	$1/2$	$1/2$	0	0
B	0	$1/3$	$1/3$	$1/3$
D	$1/3$	0	$1/3$	$1/3$
H	$1/3$	$1/3$	0	$1/3$
V	$1/3$	$1/3$	$1/3$	0
N	$1/4$	$1/4$	$1/4$	$1/4$

Table 3.1: Representation of specific DNA codes by ambiguous IUPAC codes. The non-zero entries show which of the four nucleotides (A,C,G,T) each IUPAC code is capable of representing.

PCR Failure

The first step in the Simera algorithm is to calculate how many copies of each sequence fail to amplify. These sequences are determined at the beginning of each round, and their numbers are reduced accordingly so that the inactive sequences are not referenced during the amplification step.

This will be dependent on the ratio of total sequence abundance to total combined sequence, primer and fragment abundance - i.e. the fraction of all elements present in PCR that are comprised of full sequences. In the first round of PCR there will be relatively many primers (but no fragments) and few sequences so this ratio will be small. As the rounds progress, more sequences will be generated and primers will be used up so the ratio will increase in size. It is logical to conclude that if primers and fragments are in plentiful supply then there will be fewer instances when sequences fail to bond with them to instigate amplification. This reasoning has been confirmed from results that show PCR efficiency is at its highest when amplicon quantity is at its lowest and vice versa (98).

To determine how many sequences fail to amplify completely in each round, the PCR failure rate is calculated as the parameter α and used to generate a binomial random variable for each sequence:

$$\alpha = \frac{\text{sequence abundance}}{\text{sequence abundance} + \text{fragment abundance} + \text{primer abundance}}$$

$$\text{Failures} \sim Bin(a_s, \alpha)$$

where a_s is the abundance of sequence s. The effective abundance of sequence s, X_s is the remaining number of molecules of sequence s that are available for PCR extension and

chimera formation.

$$X_s = a_s - \text{Failures}$$

Dealing With Reverse Primers

If the model is to follow the PCR process analogous then, when the simulation of a sequence splitting into two strands takes place, two differing sequences should be recorded. The first will be the original sequence of nucleotides and bind with the forward primer before commencing extension. The second sequence will be the *reverse complement* - meaning that the order of the sequence is reversed and that each nucleotide is swapped for its corresponding complementary nucleotide (A $\Leftrightarrow$ T and C $\Leftrightarrow$ G) - of the first sequence and will bind with the reverse primer.

In order to increase efficiency (and conserve memory in the implementation of the algorithm) a good shortcut is to use the reverse compeiment of the reverse primer instead of the genuine reverse primer. This means that both complementary strands for every sequence do not need to be recorded and instead only one strand is required. Binding can be simulated by attaching the forward primer and the new (reverse complement) reverse primer to opposite ends of two copies of this strand as shown in Figure 3.4.

Choosing the Best Fragments for Chimera Formation

As declared in the assumptions in Section 3.4.1, fragments will be less effective at binding with sequences than primers so, to make the model realistic, fragments must be penalised by giving more weight to the probability of a sequence binding with a primer. Some fragments will also be more adept than others at acting as primers so this must also be taken into account. This is done by comparing the last twenty nucleotides on the candidate fragment with all possible positions on the sequence. The functional part of a typical PCR primer is around twenty nucleotides long, therefore using twenty nucleotides from a fragment is a logical choice when the fragment will be acting as a primer.

The number of differences between the fragment and the sequence at each point is recorded, giving the minimum number of differences and the position at which this minimum value occurs for each candidate sequence.

In the example in Figure 3.5 it can be seen that position C gives the fewest differences between the fragment and the sequence. In this case there are zero differences compared to two

Figure 3.4: Simulated forward and reverse PCR primers. Notation referring to the direction of each primer is relative to the leading strand.

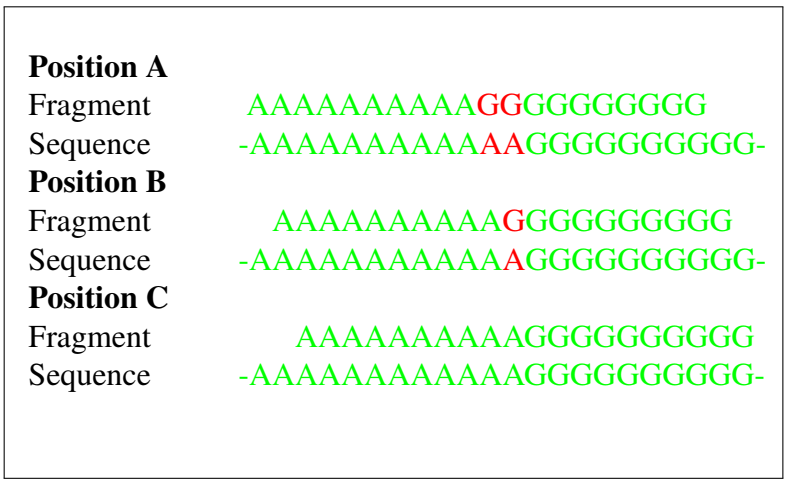

Figure 3.5: Determining the optimal position for a fragment to act as a primer. Position C is chosen because there are fewer differences.

and one in positions A and B respectively. So far, only fragments acting as forward primers have been considered. Fragments acting as reverse primers are analysed separately in the same way, except that the *first* twenty nucleotides of the fragment are compared with the sequences instead of the last twenty.

Once the optimal position and the number of differences has been found for each fragment then each fragment f can be weighted based on its suitability using the set of parameters

$$W_f = e^{-d_f}$$

$$f = 1 \dots n_{frag}$$

where d_f is the number of differences for fragment f and n_{frag} is the total number of fragments. This assigns higher weights to fragments with fewer differences, as required and all weights are forced to be between zero and one. The equation for W_f takes into account the fact that a greater quantity of energy will be required to bind fragments with a large number of differences, making it much less likely that these fragments will successfully bind.

A weight, W_p, for selecting a primer is calculated in the same way. In the case where the primer contains ambiguous IUPAC codes, a non integer number of differences may be awarded if parts of the primer result in a partial match to the sequence. For example, if the primer contains the code M then this will result in a difference of 0.5 if it is compared with either of the codes A or C (see Table 3.1). The primer is designed to be able to align well with part of the sequence so it will, typically, have very few or zero differences. It is easy to see that if there are zero differences between the primer and part of the sequence then a value of $W_p = 1$ will be calculated.

These weights, together with the set of abundances of each primer and fragment, can be used to determine which primer or fragment each sequence will bind with. Wallenius' multivariate non-central hypergeometric distribution can be used for this purpose because it models the selection of items without replacement based on their abundance and allowing unequal probabilities of selecting items of differing type, such as the primers and fragments of varying quality in this model. Selection without replacement is appropriate because when a primer or fragment binds with a sequence then it will no longer be available for use in the current round.

For each sequence, random variables are drawn from the Wallenius distribution to identify the quantity of each primer or fragment to be selected for amplification.

$$\mathbf{C} \sim Wallenius(X_s, \mathbf{A}, \mathbf{W})$$

where

$\mathbf{C} = [c_p, c_1, \ldots c_{n_{frag}}]$, $\mathbf{A} = [a_p, a_1, \ldots a_{n_{frag}}]$ and $\mathbf{W} = [W_p, W_1, \ldots W_{n_{frag}}]$.

The parameters a_p and $a_1 \ldots a_{n_{frag}}$ are the abundances of the primer and fragments respectively.

Amplification and Fragmentation

Each sequence can either bind with a fragment to form a chimera or bind with the correct primer to commence amplification. Amplification can either continue until the entire sequence has been amplified as intended or it can fail part of the way through to form a sequence fragment. When a sequence is ready for amplification it will be split into two strands, one will use the forward primer (or a forward fragment) and the other will use the reverse primer (or a reverse fragment). This means that amplification can be split into two separate processes. For each sequence to be amplified the abundance is set to zero then increased by one if the forward strand successfully amplifies and increased by one again if the reverse strand amplifies.

Consider the process to amplify forward strands - the reverse process is symmetrical and will not be described in detail. The sequences can be examined in turn. The parameter λ is used to determine whether the first nucleotide in the sequence is amplified. If it is then the second is amplified with the same probability and so on until the entire sequence is amplified. If at some point a nucleotide fails to amplify then amplification stops entirely for the sequence and the incomplete sequence is added to the pool of (forward) fragments. To model this, the primer and fragments are examined separately and binomial random variables are used for each possible fragment of sequence s. In the case of the primer,

$$Y_z \sim Bin(c_p, \lambda)$$

$$z = (l_p + 1) \ldots (l_s - 1)$$

where l_s and l_p are the length of the sequence and the primer respectively. The new fragment is created by joining together the l_p nucleotides of the primer with nucleotides in positions $(l_p + 1)$ to z in sequence s. Y_z copies of the fragment are added to the pool of fragments and c_p is reduced by Y_z. The process is then repeated for each (old) fragment, f, in place of the primer, substituting c_f and l_f for c_p and l_p, respectively.

For each sequence there are $(l_s - l_p)$ possible fragments. This is the case because amplification can fail at position $(l_p + 1)$ through to position l_s, giving $(l_s - l_p)$ possible failure points. The integer z is the same as the length of the fragment created.

The PCR round is now complete and a new round can commence.

3.4.2 Implementation

The Simera algorithm was implemented using **C++** code. This implementation makes use of the *randomc* and *stocc* libraries (99) which provide the random number generator and probability distributions necessary to implement the algorithm. The latter of these libraries required slight modification to enable compatibility.

The program requires as input the sequences to be amplified and their initial abundances, the primer pair, the number of rounds of PCR to be simulated, the number of reads to be sampled post-simulation and the value of the parameter λ.

Pre-processing and Formatting

The input files and parameters must be in the correct format for the Simera program to function correctly. The number of rounds of PCR to be simulated, the number of reads to be sampled post-simulation and the value of the parameter λ can be supplied as command line input and the primer pair can be supplied as a fasta file. The sequences to be amplified must also be in fasta format with each sequence having a unique name containing the sequence's abundance as the final part of this name. The sequences themselves must be truncated so that only the target regions to be amplified, flanked on either side by the two primer-compatible regions, are present.

3.4.3 Calibration

To determine the value of the parameter λ, simulated data were compared with the experimental data described in Section 2.3. To mimic this experiment, the good sequences (as detected by Perseus) from the experiments containing 12, 24 and 48 closely and distantly related nematode species were used as input for 35 simulated rounds of PCR - the same number of rounds as the original experiment. The same number of reads produced for each experiment were sampled from the output of each corresponding simulation and the number of chimeras produced in each case were recorded. Different values for λ were tried, each experiment was repeated five times and the value that gave the closest match between the

experimental data and the simulated data was found to be $\lambda \approx 5 \times 10^{-6}$ as can be seen in Figure 3.6.

This value for λ can be considered accurate for simulations of PCR under the same conditions as those used to generate the experimental data. To simulate PCR with different conditions, different values for λ may be more appropriate.

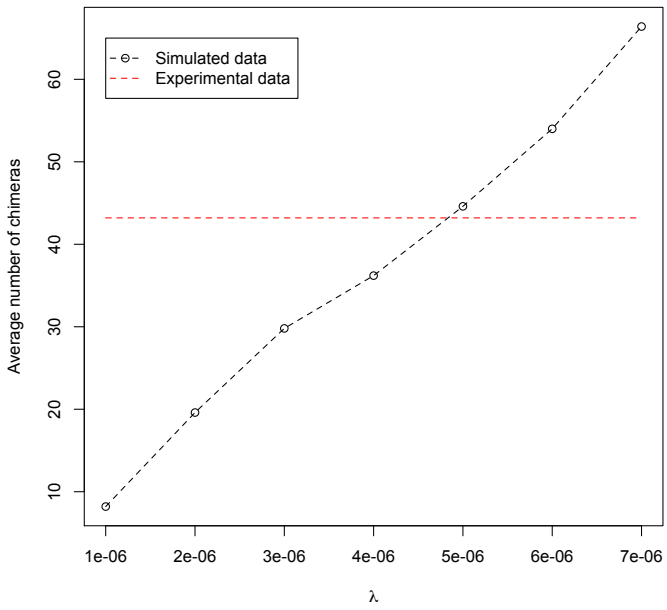

Figure 3.6: Number of chimeras simulated using the Simera algorithm for different values of λ. The parameter λ is the failure rate, during the extension step, at each nucleotide position on the query sequence. 35 rounds of PCR were simulated using the good sequences (as detected by Perseus) from pooled experiments on 12, 24 and 48 closely and distantly related nematodes as input.

3.4.4 Results

In order to assess the performance of the model, simulated data were again compared with the experimental data described in Section 2.3. The good sequences (as detected by Perseus) from each of the closely and distantly related pooled nematode experiments were used as input for 35 simulated rounds of PCR. True break points and parents are available as output from the simulation software, however to compare the simulated data with realistic data it was necessary to find the break points and parents in the same way as the original experiment.

As with the analysis described in Section 2.3, the output from Perseus returned most likely break point for each chimera based on its two identified parent sequences. These break points were standardised for the whole dataset by forming a four-way alignment of each chimera,

its two parents and the *C. elegans* reference sequence using ClustalX (16). The position of each break point on the reference sequence was recorded to give a standardised break point. The frequency of each standardised break point could then be recorded to assess which regions of a sequence were most susceptible to chimera formation.

Break point frequencies from the experimental and simulated data are shown in conjunction with the equivalent results for the second algorithm in Section 3.5.4 (Figure 3.12) and their distributions appear to be similar. A Kolmogorov-Smirnov test, adapted for use with discrete distributions in the *dgof* **R** package (100), was performed and yielded a p-value of 0.607, indicating that there was no evidence that the two sets of data were drawn from distinct distributions. In addition to this, the two sets of break point frequencies have a Pearson's correlation coefficient of 0.735. It can be inferred from these results that the simulated data are distributed similarly to the experimental data.

3.5 Model 2

3.5.1 How Can Model 1 be Improved?

It has been shown in Section 3.4 that the first PCR model is a faithful model of the PCR process which can accurately simulate the generation of realistic chimeras. The main negative issue with the model is that the implementations of it run too slowly to be useful for studies involving medium-sized to large datasets.

Ways of generalising and adapting the model must, therefore, be sought in order to increase the speed of simulations without significantly reducing the accuracy and reliability of the output. This is achieved in this section by taking a more abstract approach which involves creating a pool of the most likely chimeras prior to the main body of the algorithm being executed. All simulated chimeras may now only come from this pool and this, in turn, means that individual fragments no longer need to be recorded. Instead, only the overall number of fragments is required.

3.5.2 Model Outline

The updated algorithm for Model 2 is named *Simera 2*. The two parts of the Simera 2 algorithm are described in detail in the remainder of this section. The complete algorithm is presented in Figure 3.7 and visualised in Figures 3.8 and 3.9.

Assumptions

In this model chimeras are still formed in the same way, the difference is that rarer chimeras will now be ignored. Therefore, in addition to the assumptions made for Model 1, it is assumed that rare chimeras will be generated in low enough abundances during PCR that they will not be selected when reads are sampled during sequencing.

Input and Parameters

Most of the input for the second algorithm is the same as the first:

1. A list of initial sequences and their initial abundances.
2. Forward and reverse primers and their initial abundances.
3. λ – The rate of failure at each nucleotide on a sequence.

- Set initial pool of sequences and abundances, total abundance is a_{tot}.
- Set empty pool of chimeras of size c_{tot}.
- Set primer sequences and initial primer abundances, a_p.
- Set λ.
- Set number of PCR rounds.
- Set number of chimeras to generate, c_{gen}.
- Repeat c_{gen} times:
  - Select two random sequences, s_i and s_j of length l_i and l_j $(l_i < l_j)$.
  - Generate a random break point, b, on s_i.
  - Form chimera from first b bases of s_i and the last $l_j - b$ bases of s_j.
  - Calculate probability of this chimera forming.
  - If probability is in the highest c_{tot} probabilities, add to pool of chimeras.
- For each round of PCR, r:
  - Set potential amplification pool:
    - new a_{tot} = old $a_{tot} \times 2$.
  - Reduce a_{tot} by PCR failures:
    - Efficiency $= \frac{\text{primer abundance+frag abundance}}{\text{prim abundance+frag abundance+seq abundance.}}$
    - Fail rate $= 1-$ Efficiency.
    - Failures = Binomial(Fail rate, a_{tot}).
  - Reduce a_{tot} by fragments formed:
    - Prob(fragment) $= 1 - [1 - \lambda^{(\text{seq length}-\text{primer length})}]$.
  - Randomise chimeras and sequences:
    - $\beta = \frac{\text{mean primer weight}\times\text{primer abundance}}{\text{mean prim weight}\times\text{prim abund+mean frag weight}\times\text{frag abund.}}$
    - sequences = Binomial(β, a_{tot}).
  - Increase individual sequence abundances:
    - Hypergeometric (sequences, old sequence abundance vector).
  - Select chimeras:
    - Multinomial (chimeras, chimera prob vector).
  - Add selected chimeras to sequence pool.
  - Reduce primer abundance by number of sequences formed.
  - Increase fragment abundance by number of fragments formed.
  - Reduce fragment abundance by number of chimeras formed.
- End of algorithm.

Figure 3.7: Simera 2 algorithm for Model 2. λ is the rate of failure during PCR extension at each nucleotide point on a sequence.

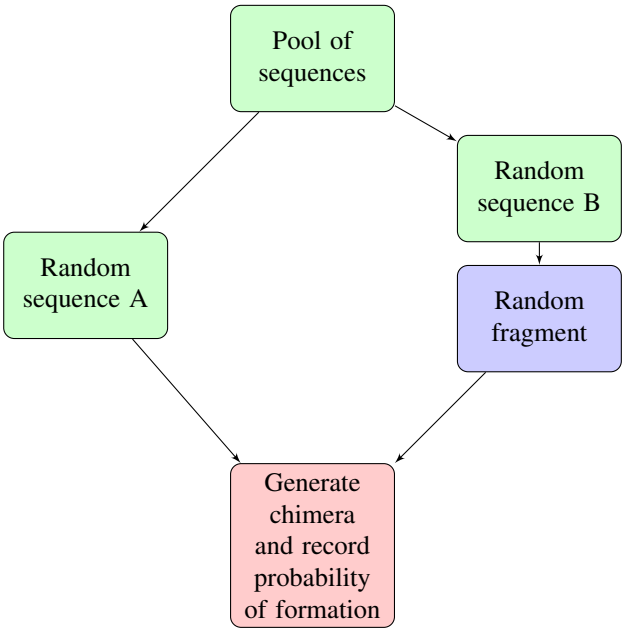

Figure 3.8: PCR simulation using Model 2: Chimera formation step. This step is to be repeated until the desired number of chimeras is reached.

4. A list of potential chimeras - a specified number of chimeras are to be recorded for use later on in the algorithm. The pool is empty initially.

Chimera Formation Step

The Simera 2 algorithm comprises of two steps. The first of these involves constructing all possible chimeras and calculating the probability of each chimera forming. At a later stage in the algorithm, the probabilities associated with the individual chimeras will be used to select a chimera at random when one is created. This removes the need to generate a new chimera every time one is formed and, it is hoped, will not impact on the realism of the first model.

To generate all possible chimeras, all possible fragments are aligned with all sequences and the best chimera (fewest mismatches) is found, as described in Section 3.4.1. The probability of a fragment of length l forming is

$$\text{Prob} = \lambda(1-\lambda)^k$$

where $k = l - p$ and p is the length of the primer used to form the fragment. The integer, k, is the same value as the number of successful nucleotide extensions prior to failure.

This probability is then combined with the relative abundance of each sequence and the number of mismatches between the fragment and the sequence to calculate the probability

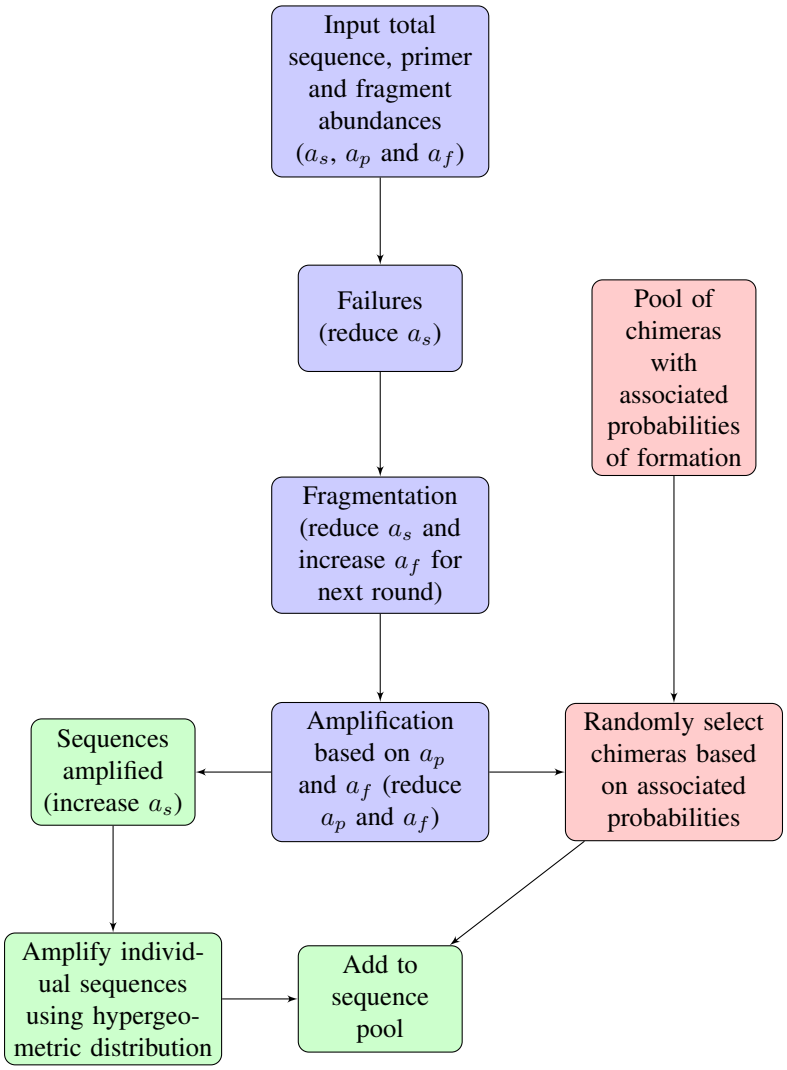

Figure 3.9: PCR simulation using Model 2: PCR step. Repeat this step for the desired number of PCR rounds.

of each chimera forming,

$$\text{Prob} = \lambda(1-\lambda)^k a_i a_j \text{e}^{-m}$$

where a_i and a_j are the relative abundances of the two sequences and m is the number of mismatches.

A specified number of the most probable chimeras are recorded for later use. For larger datasets, generating all possible chimeras will be too computationally intensive and, instead, a predefined large number of chimeras may be generated randomly and the most probable chimeras are selected from these.

PCR Step

The following PCR step is intended to approximate the first Simera algorithm outlined in Section 3.4, it is to be repeated for the specified number of rounds.

PCR failures are calculated using exactly the same method as in the original model, except they can be calculated for all sequences together rather than each sequence separately:

$$\text{Failures} \sim Bin(a_t, \alpha)$$

where a_t is the total sequence abundance. The probability of a sequence fragmenting is

$$\text{Prob}_{frag} = 1 - (1-\lambda)^k$$

which is just one minus the probability of the sequence amplifying successfully. This probability can then be used to calculate the number of fragments created for each sequence in the current round:

$$\text{Fragments} \sim Bin(a_s, \text{Prob}_{frag})$$

where a_s is the abundance of sequence s. This works the same way as fragmentation in the first Simera algorithm but here only the number of fragments is recorded instead of each fragment being recorded individually. This method avoids the need to use Wallenius' distribution to select individual fragments based on their weightings.

The number of sequences available for amplification, a_t, is reduced by the number of failures and fragmented sequences.

The number of successfully amplified sequences is determined using the parameter β, where

$$\beta = \frac{\text{mean primer weight} \times \text{primer abundance}}{(\text{mean prim weight} \times \text{prim abundance}) + (\text{mean frag weight} \times \text{frag abundance})};$$

$$\text{Amplified sequences} \sim Bin(a_t, \beta).$$

The number of individual sequences amplified is then found using a hypergeometric random variable, as follows:

$$\mathbf{S} \sim Hypergeometric(\text{Amplified sequences}, \mathbf{A})$$

where **A** is the vector of individual sequence abundances. This amplifies the sequences all at once, compared with the first Simera algorithm which amplifies each sequence separately using successive binomial random variables. Accuracy is lost because mean fragment weights are used rather than the exact values.

Any remaining sequences are used to generate chimeras and this stage is where the two algorithms differ the most. In the Simera 2 algorithm, the chimeras are chosen from the prepared pool using a multinomial distribution rather than being created from a fragment pool when required. This is much quicker.

$$\mathbf{C} \sim Multinomial(\text{Remaining sequences}, \mathbf{P})$$

where **P** is the vector of the probabilities of formation for each chimera. Generated chimeras are then added to the pool of sequences for the next PCR round.

3.5.3 Implementation

The Simera 2 algorithm was implemented using **C++** and the program has the same input requirements and dependencies as the original Simera program.

3.5.4 Results

The implementation of the Simera 2 algorithm has been shown to be able to handle large, realistic datasets and it is used for this purpose in the following chapter. To verify that this algorithm is a good approximation of Simera, both algorithm implementations were used to simulate chimeras for the same datasets - the closely and distantly related nematode pools which were also used in Section 3.4.4 - with the same input parameters (35 rounds of PCR and $\lambda = 5 \times 10^{-6}$). It was not possible to compare the models using larger datasets because

of the inhibitive speed of the original Simera implementation. However, it was theorised that if the models work comparably on smaller datasets then the same should be the case for the larger datasets analysed in the next chapter.

When the simulated output from the Simera 2 algorithm was subsampled, using the same sample size as that used for the original Simera data, the average number of chimeras produced was 42.2, compared with 44.6 for the original Simera data and 43.2 for the real experimental data.

To compare the type of chimeras which were formed, the break points were again observed. On this occasion there was no need to use the method involving Perseus and the *C. elegans* reference sequence because all break points were available as output from the simulation software. The break points of all chimeras generated in each of the simulated experiments were compared, and the distributions of these are shown in Figure 3.10.

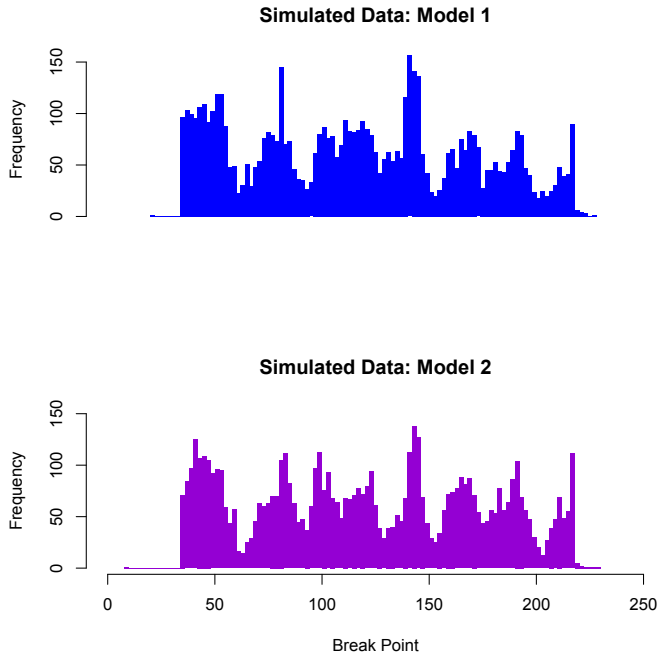

Figure 3.10: Break point frequencies for simulated data comparing results from the Simera and Simera 2 algorithms. For each algorithm, 35 rounds of PCR were simulated using the good sequences (as identified by Perseus) from pooled experiments on 12, 24 and 48 closely and distantly related nematodes as input. Break points are returned as output from the Simera and Simera 2 implementations.

The two distributions appear very alike, suggesting that both models produce the same chimeras and that Model 2 is an excellent approximation of Model 1. To verify these assertions, a Kolmogorov-Smirnov test, again using the *dgof* package in **R**, was performed on the two sets of frequency data. A p-value of 0.970 was returned, indicating that there was no

evidence that the two datasets were drawn from distinct distributions. Additionally, the two datasets of break point frequencies are closely correlated, as can be seen in Figure 3.11, with a Pearson's correlation coefficient of 0.914. These results provide strong evidence that the two algorithms generate very similar output when provided with identical input.

In addition to the true break points provided by the software, the break points found by aligning each chimera, its parents (as detected by Perseus) and the *C. elegans* reference sequence were also recorded and again compared with those from the experimental data. The break point frequencies are shown in Figure 3.12 and they appear to be distributed similarly to the experimental break points as well as the break points generated using the Simera 1 algorithm. A two sample Kolmogorov-Smirnov test between the Simera 2 break points and the experimental break points returned a p-value of 0.592, meaning that there was no evidence to suggest that the two samples were differently distributed, and the two samples were positively correlated with a Pearson's correlation coefficient of 0.682.

The quality of chimeras generated with the Simera algorithms was compared with that of those generated using the existing PCR simulator, Grinder 0.5.3. Two different methods of chimera generation were used. The first was Grinder's default method which simply creates chimeras based on a random break point, the second method applies the same technique as used by CHSIM which requires both parents to share an identical k-mer of length 10. In order to specify the required k-mer length, Grinder must be supplied with the input parameter, 'ck', so the first method has $ck = 0$ and the second has $ck = 10$.

Break points for the chimeras generated by Grinder were found using the same method as was used to find the break points of those generated using the Simera algorithms, and the distributions are shown in Figure 3.13. As expected, the distribution when $ck = 0$ is fairly flat and bears little similarity to the distribution of the experimental break points. The distribution when $ck = 10$ seems more realistic with some peaks and troughs appearing in the same regions. However, there is an excessively large number of break points over 200 and the remainder of the peaks are not as high as their experimental counterparts. Overall, to the naked eye, the distribution does not seem as realistic as those generated from the Simera algorithms.

All simulated sets of break points were compared with the set of experimental break points and the Kolmogorov-Smirnov p-values, along with the Pearson's correlation coefficients, are displayed in Table 3.2. The p-value of 0.471 and correlation coefficient of 0.520 give no indication that chimera break points generated using Grinder with $ck = 10$ are distributed differently from the experimental data but these numbers are lower than those found using

the Simera algorithms which means that there can be greater confidence that the Simera-generated chimera break points share the experimental distribution. The very low p-value shown supplies very strong evidence that the break points generated using Grinder with $ck = 0$ are not distributed in the same way as the experimental break points.

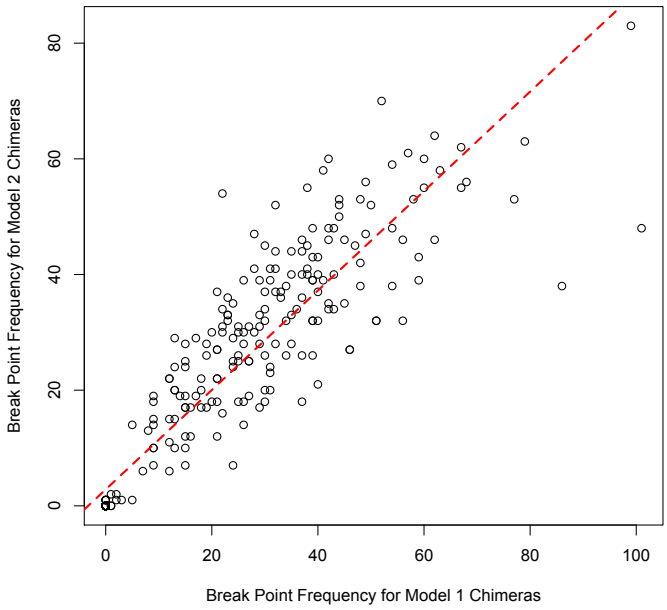

Figure 3.11: Break point frequencies for simulated data generated using the Simera algorithm plotted against the same data generated using the Simera 2 algorithm. 35 rounds of PCR were simulated using the good sequences (as identified by Perseus) from pooled experiments on 12, 24 and 48 closely and distantly related nematodes as input. Break points are returned as output from the Simera and Simera 2 implementations.

Simulation Method	K-S Test p-value	Pearson's Correlation
Simera 1	0.607	0.735
Simera 2	0.592	0.682
Grinder (ck=0)	0.000005	0.289
Grinder (ck=10)	0.471	0.520

Table 3.2: Kolmogorov-Smirnov test p-values and Pearson's correlation coefficients returned when various sets of simulated break points were compared with experimental break points.

To investigate sequence similarity between simulated and real data, the chimeras generated from the closely and distantly related experiments were compared with the chimeras generated from the corresponding simulations using both algorithms. Using the good sequences which were detected by using Perseus on the experimental output, each simulation was repeated five times and the generated chimeras were pooled to create reference datasets for each simulated experiment. The reference datasets were subsampled so that both Simera and Simera 2 reference datasets were the same size for each experiment. USEARCH was used

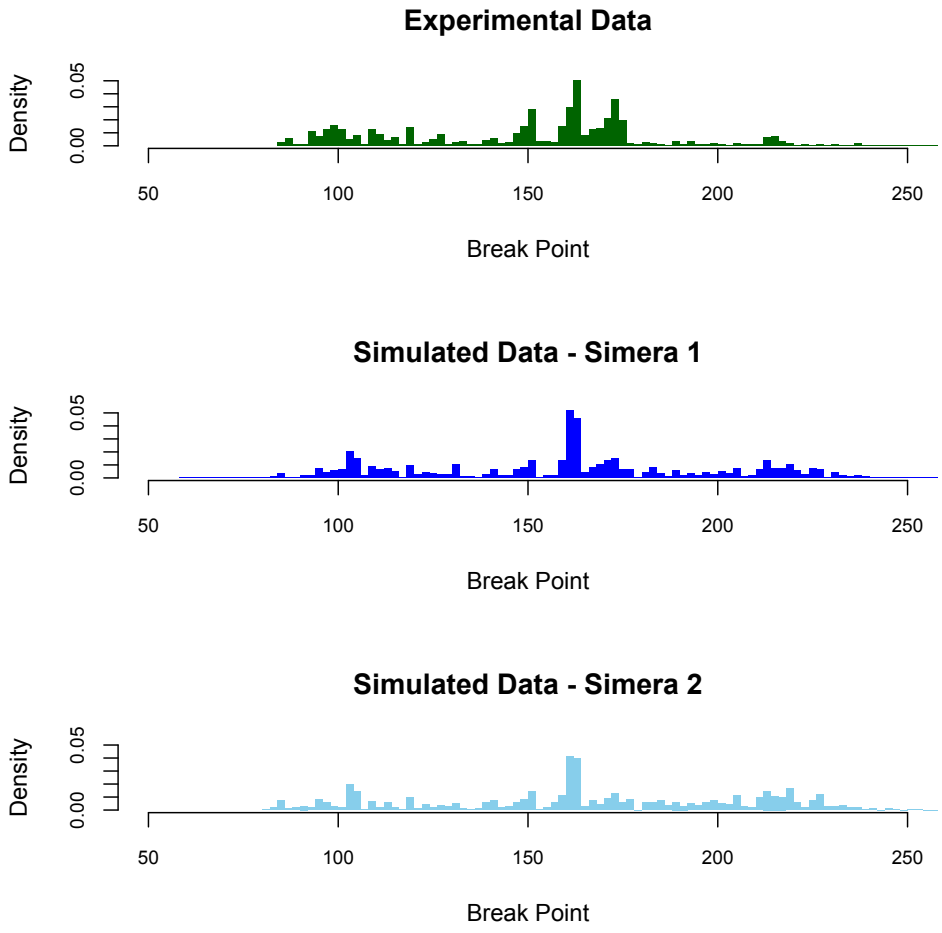

Figure 3.12: Break points of chimeras generated from the Simera and Simera 2 algorithms compared with those from pooled experiments on 12, 24 and 48 closely and distantly related nematodes. For the simulated chimeras, 35 rounds of PCR were simulated using the good sequences (as identified by Perseus) from the pooled experiments as input. In all three cases a four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded.

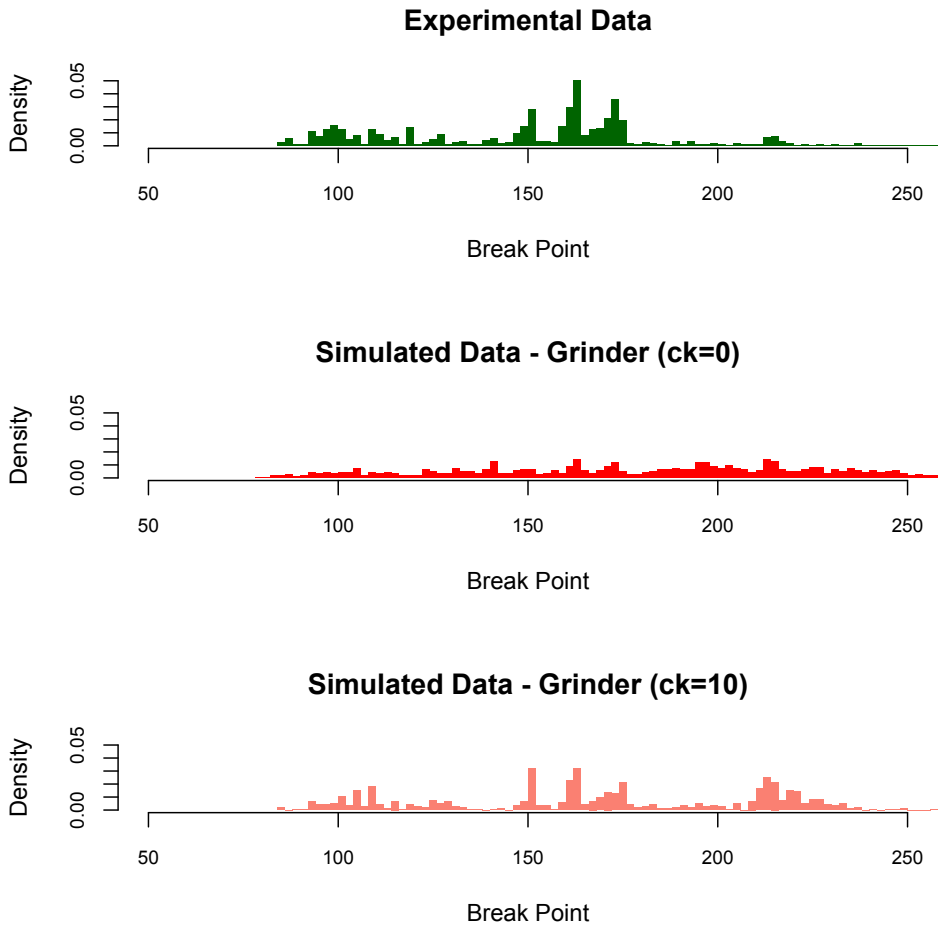

Figure 3.13: Break points of chimeras generated from Grinder, using both $ck = 0$ and $ck = 10$, compared with those from pooled experiments on 12, 24 and 48 closely and distantly related nematodes. For the simulated chimeras, 35 rounds of PCR were simulated using the good sequences (as identified by Perseus) from the pooled experiments as input. In all three cases a four way alignment was formed, using ClustalX, between each chimera, its two parents (as identified by Perseus) and the *C. elegans* reference sequence. The number of break points (as identified by Perseus) at each point on the alignment were recorded.

to find the closest matches between chimeras in the query dataset and those in the reference datasets and the number of exact matches (100% similarity) and matches with greater than 99% similarity were recorded. The two simulated datasets were compared against each other in the same way to see how many chimeras were present in both.

Figure 3.14 shows the results from this analysis. The output from Simera contained slightly more identical matches to the experimental chimeras than the Simera 2 output (31.5% versus 28.8%) and also slightly more near ($>$ 99%) matches (59.7% versus 53.2%). The two simulations were shown to be producing similar chimeras with approximately 80% of the chimeras produced using Simera 2 closely matching ($>$ 99%) those produced by Simera.

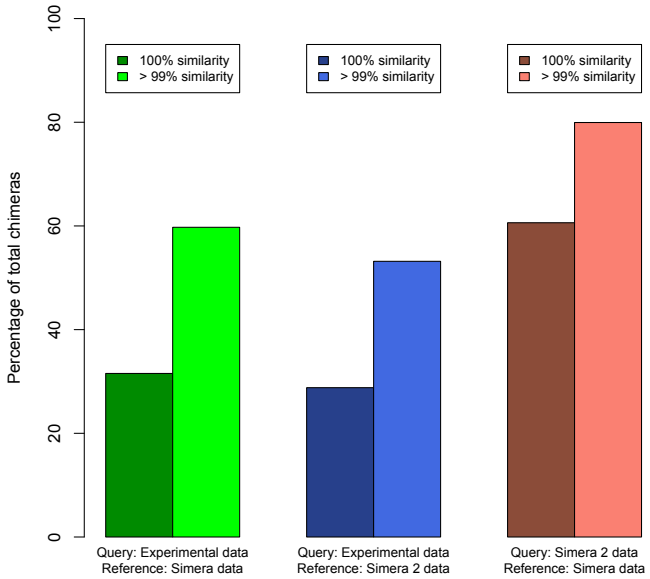

Figure 3.14: Sequence similarities (using USEARCH) when comparing experimental chimeras against datasets of simulated chimeras and when comparing datasets generated using the two different Simera algorithms.

Species Relatedness	Number of Species	Number of Chimeras in Reference Databases	Expected 100% Matches	100% Matches (Simera)	100% Matches (Simera 2)
Close	12	4744	18.0%	20.0%	20.0%
Close	24	7309	6.6%	38.1%	42.9%
Close	48	15822	3.5%	18.4%	15.8%
Distant	12	2442	9.3%	44.4%	38.9%
Distant	24	5093	4.6%	35.6%	26.4%
Distant	48	11018	2.4%	32.7%	28.8%

Table 3.3: Expected 100% matches for experimental chimeras versus actual 100% matches when compared with reference datasets of chimeras generated using the Simera and Simera 2 algorithms. USEARCH was used to determine percentage similarity between sequences.

The number of exact matches between experimental chimeras and those simulated for the reference sets was far greater than would be expected for randomly generated chimeras. Examining, for example, the experiment on 24 closely related nematodes, the number of different potential chimeras can be calculated. Each of the 24 input sequences is 220 base pairs in length, meaning that there are 200 potential break points on each sequence. This results in 200 different potential fragments for each sequence which, multiplying by 24, makes 4800 total potential fragments. Each of these fragments can form one chimera when paired with any of the 23 other sequences, so there are $23 \times 4800 = 110400$ possible chimeras resulting from this dataset. As there were only 7309 chimeras generated for the reference datasets used for this experiment then, under a random uniform model of chimera generation, the expected proportion of the experimental chimeras appearing in the reference datasets is $7309/110400 = 0.066$ or 6.6%. This result contrasts with the actual percentage of matches for this experiment which were 38.1% for the reference dataset of chimeras generated using the Simera algorithm and 42.9% for the reference dataset of chimeras generated from the Simera 2 algorithm.

Table 3.3 shows the expected exact matches for each of the six experiments. In every instance these are significantly lower than the actual exact matches which is strong evidence that the simulated chimeras are more realistic than uniform randomly generated chimeras. Expected matches are lower for experiments with a higher number of input sequences because there are more potential chimeras for these. This analysis does not consider chimeras generated from other chimeras but the inclusion of these would further increase the number of different potential chimeras and, therefore, further decrease the amount of expected matches.

3.6 Discussion

The model presented in Section 3.4 provides an accurate representation of PCR. It has few parameters and assumptions and the output is shown to reflect real experimental results. The drawbacks of this model are related to speed limitations associated with the implementation of the algorithm and its usage is restricted to very small datasets which mean that it can't be used for the majority of analyses.

A second model is presented in Section 3.5, the algorithm of which solves the problems associated with Model 1. Furthermore, the results obtained from simulations utilising the second model's methodology show no indication of being any less accurate than those obtained from simulations involving the first model's methodology.

The abundance and nucleotide composition of chimeras generated, both *in vitro* and *in silico*, can vary somewhat even when PCR is performed under the same conditions on identical samples. This is, presumably, caused by the random occurrences of PCR extension failure and makes it difficult to determine the degree of realism present in the data output from the simulations. However, results show that around a quarter of the chimeras found in the experimental data involving pooled nematode samples were reproduced perfectly during the simulated experiments and around a half were reproduced at a level of better than 99% similarity. Because of the potential for PCR and sequencing noise, it is reasonable to suggest that some of the closely matching chimeras were, in fact, exact matches.

The fact that many chimeras were reproduced exactly is encouraging, as are the results showing that the chimera break points are distributed similarly in experimental and simulated datasets. This is evidence that the chimeras are being generated in the same way, i.e. that the same 'type' of chimeras are being produced, even if their nucleotide composition is subject to natural variance. Furthermore, the distributions of break points on chimeras generated using the Simera algorithms compare favourably to the distributions of those on chimeras generated using Grinder. This is evidence that Simera generates more realistic chimeras than the best existing PCR simulators.

Section 3.1.1 discusses some existing PCR simulation software and notes that most available tools involve the selection of amplicons from a reference database by matching primer sequences to areas of similarity on the reference sequences. Because both Simera algorithms require ready-made amplicons as input, the models presented in this chapter work best when used in conjunction with existing software. For example, amplicons can be selected from a reference database using Primer Prospector and these amplicons can then be used as input for one of the Simera algorithms. This procedure is followed, and explained in more detail, in Chapter 4 for the generation of *in silico* datasets.

Overall, it can be concluded that both models presented in this chapter can be used to produce realistic simulated PCR output, particularly with respect to the chimeras generated during the process. In addition, the Simera 2 algorithm can be implemented sufficiently well to allow these simulations to be carried out on large, realistic datasets.

Chapter 4

Analysis of *In Silico* Datasets

4.1 Introduction

The previous chapter describes the development and implementation of the PCR simulation algorithm. One of the most useful applications of the simulation software is its utilisation in the generation of *in silico* datasets representing real next generation sequencing data. The same simulations can be replicated numerous times to find the level of variance in the output. Different input parameters can also be changed to find the PCR setups corresponding to the best performance in reconciling the noisy dataset with the input dataset. Results should give an indication of the reliability of the information found from microbial community analysis.

The analysis of these *in silico* datasets is most useful for measuring the performance of chimera removal software, as it is simple to check whether the known chimeras in the data are detected or not. Realistic *in silico* datasets with identifiable chimeras will provide a more accurate performance analysis of chimera removal software than the previous testing methods which used mock community datasets.

4.2 Methods

Various methods were used to generate *in silico* datasets but not all methods were used for every dataset. This section presents a general overview of all methods used. In the results section (Section 4.3) specific information is included about how the different datasets that were analysed were generated.

4.2.1 Clustering

All OTU Clustering was performed using UCLUST (97) integrated in the QIIME environment (21). UCLUST uses a centroid-based medium to high-identity clustering algorithm where sequences are added to clusters based on their similarities to the *centroid sequence*. This centroid sequence is the sequence that most closely represents all members of the cluster.

Unless otherwise stated, the default UCLUST settings of 97% OTUs, *de novo* cluster identification, pre-filtering of identical sequences and pre-sorting of sequences by abundance were applied. In some cases 99% OTUs and cluster identification using a reference database were used instead.

4.2.2 Selecting Primers and Amplicons

Suitable virtual primers for the chosen data were selected, these were the forward primer:

515F (5′-GTGNCAGCMGCCGCGGTAA-3′)

and the reverse primer:

806R (5′-GGACTACHVGGGTWTCTAAT5′-).

These primers are often used to highlight regions of the 16S gene. They were tested using Primer Prospector (94) on the data to be used for creating the *in silico* dataset and the output was analysed to ensure that they were effective primers. The amplicons that were used in the various *in silico* datasets were selected using these primers in Primer Prospector.

4.2.3 Adding an Abundance Distribution

Abundance distributions were generated using the inbuilt random variable generating functions in **R**. The required number of random variables (matching the number of sequences in the dataset) and the chosen values for the distribution parameters were used as input.

4.2.4 Simulating PCR

PCR was simulated using the Simera 2 software which runs the algorithm described in Chapter 3. Simera 2 requires that the number of PCR rounds and the number of reads to be sam-

pled are used as input along with the fasta file that represents the community to be sequenced.

4.2.5 Simulating PCR Single-Base Errors

PCR single-base errors were randomly added to fasta files when required. The probabilities of each error occurring are shown in Figure 4.1 and these were found during the development of AmpliconNoise (18).

Nucleotide	A	C	G	T
A	0.9995	7.2×10^{-6}	5.1×10^{-4}	7.7×10^{-6}
C	1.1×10^{-5}	0.9996	2.1×10^{-6}	4.1×10^{-4}
G	3.5×10^{-4}	3.2×10^{-6}	0.9996	2.1×10^{-5}
T	9.0×10^{-6}	5.7×10^{-4}	1.4×10^{-5}	0.9994

Table 4.1: Probabilities of single base errors based on data from mock communities (18). Rows are the true nucleotides and columns are those erroneously observed.

4.2.6 Simulating Sequencing Noise

Flowsim (101) was used to simulate sequencing noise on fasta files when required, generating simulated flowgram files. Flowgram files were converted back to fasta format using scripts available in QIIME.

4.2.7 Noise Removal

The AmpliconNoise pipeline was used to remove simulated noise when required, as described in Chapter 1.

4.2.8 Chimera Detection

UCHIME v4.240 has been shown to have approximately the same success rate at detecting chimeras as Perseus (18) but runs much quicker. Therefore, to process the high volume of chimera checking required for this analysis, UCHIME was chosen and executed with the default input values.

UCHIME can either be used in "*de novo* mode" or "reference mode". The first method uses only the subject dataset to identify potential parent sequences and, consequently, chimeras whereas the second method makes use of a reference database in order to identify known non-chimeric sequences. Both methods were investigated and the results from each compared with each other. Where appropriate, one of the two recommended reference databases

for 16S data was used. These databases are the ChimeraSlayer (102) reference database and the RDP classifier training database (v9) (103).

Additional analysis was carried out on a subset of the available simulated data to compare the results obtained from UCHIME in *de novo* mode against those obtained from Perseus. Similarly UCHIME in reference mode was tested against ChimeraSlayer which also uses a reference database.

4.2.9 Generating Datasets

Two different databases were selected for this study - Greengenes (104) and Silva (105).

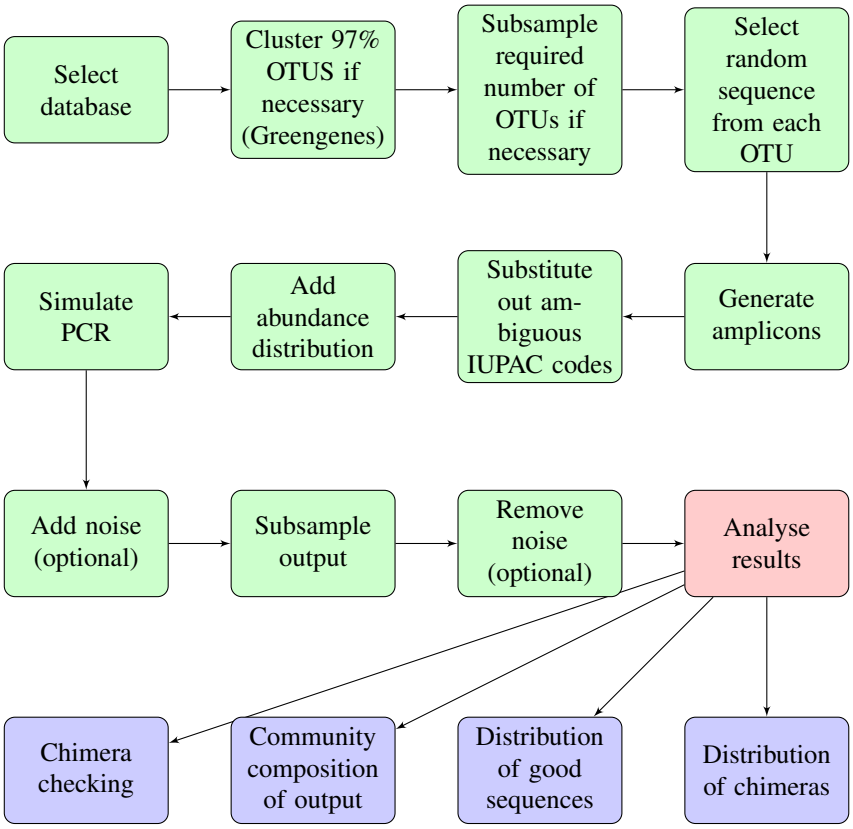

Figure 4.1: Steps followed for the generation and analysis of *in silico* datasets.

Dataset 1 - Greengenes

The original database contained sequences describing the 16S gene of 381226 individual bacteria and archaea specimens in fasta format. Some of these sequences were taken from organisms of the same species so, to ensure that the *in silico* mock community contained no duplicates, the sequences were clustered into 97% OTUs using UClust (97).

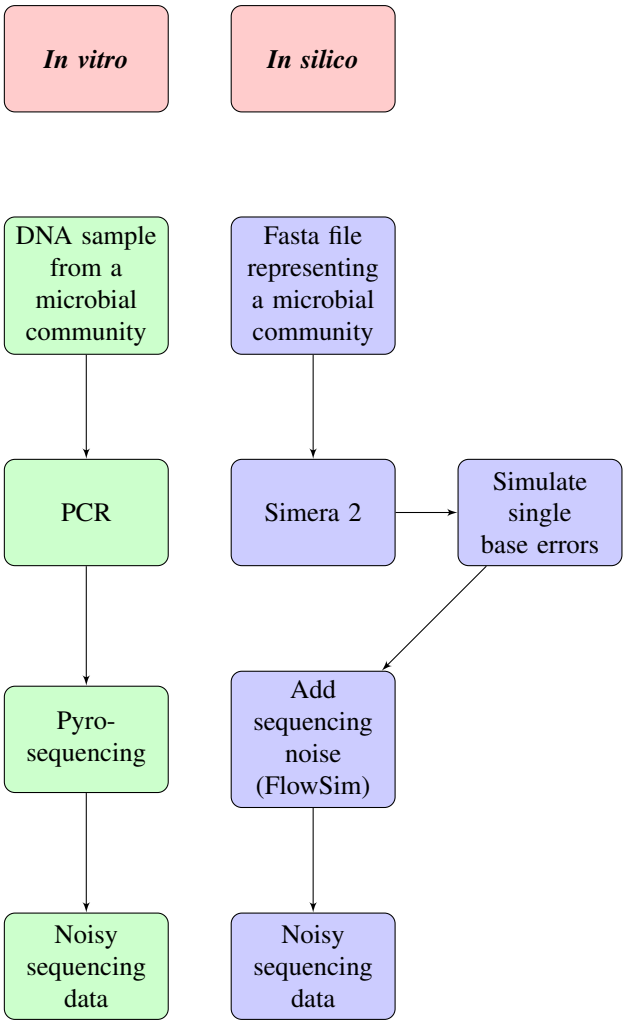

Figure 4.2: *In vitro* versus *in silico* datasets. The flow chart shows that steps followed in this chapter for creating *in silico* datasets are comparable to the amplification and sequencing steps used in laboratory based analysis. Boxes on the same horizontal level in each chain are analogous.

One sequence was chosen at random from each cluster to select the organisms present in the dataset. The region for amplification was selected using a primer pair of 515f and 806r; amplicons were extracted from the sequences using Primer Prospector. Any ambiguous IUPAC DNA base codes $[R, Y, S, W, K, M, B, D, H, V, N]$ in the fasta file were randomly replaced by an appropriate IUPAC code relating to a specific nucleotide $[A, C, G, T]$ with the probability of specific nucleotide replacement shown in Table 3.1 in the previous chapter. This was to allow the sequences to represent real data and also for compatibility with the simulation software.

Following these processes, the dataset contained 7870 sequences and was ready for a distribution of abundances to be added prior to simulated PCR amplification and noise addition. This resulting dataset is designed to represent a realistic microbial community that can be used for the purposes of testing existing noise removal software and for optimising PCR conditions to increase accuracy.

Dataset 2 - Silva

The Silva dataset is already arranged by species so clustering was not required. 20000 sequences were selected randomly and, as with the Greengenes dataset, the 515f and 806r primer pair were used to select the amplification region in Primer Prospector. Any duplicate sequences in the amplified region were removed and, as with the Greengenes dataset, ambiguous IUPAC codes were substituted for specific codes.

8000 sequences were randomly selected from the resulting dataset to create the final dataset. This number was chosen for three reasons: it is high enough to model a real life dataset; it is low enough to allow the simulation software to run effectively; it is of a similar size to the Greengenes dataset, allowing meaningful comparisons to be made between the two.

Following these processes the dataset was prepared for the addition of an abundance distribution and simulated noise.

4.2.10 Choice of Abundance Distribution

The Log-normal Distribution

The log-normal distribution is a continuous probability distribution closely related to the Gaussian, or normal, distribution. The natural logarithm of a log-normal random variable

is normally distributed with mean μ and variance σ^2. Thus if X is log-normally distributed such that $X \sim \ln N(\mu, \sigma^2)$ and $Y = e^X$ then $Y \sim N(\mu, \sigma^2)$. The log-normal distribution has probability distribution function,

$$\text{Prob}(X = x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}. \tag{4.1}$$

Where $x > 0$.

Log-normal Distribution as a Model for Species Abundance in Microbial Communities

The log-normal distribution has frequently been discussed as a model for species abundance, both in bacteria and other organisms (106) (107) (108) (109) (110).

The data used in this chapter must be discretely distributed. Although the log-normal distribution is a continuous distribution, it can still be used because the range of abundances and the number of species represented in each dataset are both deemed high enough for a continuous distribution to be a valid approximation.

Fitting a Log-normal Distribution to Experimental Data

Maximum likelihood estimates (MLEs) of the log-normal parameters for a given dataset can be found by using the experimental data and the log-normal probability distribution function (4.1). From (4.1) the likelihood function can be calculated as the product of the probabilities of each of the n observations in the dataset ($x_1 \ldots x_n$) occurring:

$$L(\mu, \sigma^2 | x_1 \ldots x_n) = \prod_{i=1}^{n} \left[\frac{1}{x_i \sigma\sqrt{2\pi}} e^{-\frac{(\ln x_i - \mu)^2}{2\sigma^2}} \right].$$

It is possible to express this function as a product of probabilities because the observations are assumed to be independent and identically distributed (i.i.d.)

Maximising the likelihood function results in finding estimates of the parameters μ and σ^2 which are most likely to have yielded the given data. In practice it is easier to maximise the log-likelihood function (which is simply the natural logarithm of the likelihood function). This has the same effect because as a variable increases, its logarithm also increases and, therefore, both functions will be maximised with the same values of μ and σ^2. The log-likelihood function for the log-normal distribution is shown below.

$$l(\mu, \sigma^2 | x_1 \dots x_n) = \ln(L) = \ln\left(\prod_{i=1}^{n}\left[\frac{1}{x_i\sigma\sqrt{2\pi}}e^{-\frac{(\ln x_i-\mu)^2}{2\sigma^2}}\right]\right),$$

$$l(\mu, \sigma^2 | x_1 \dots x_n) = -\sum_{i=1}^{n}\left[\ln(x_i\sqrt{2\pi}) + \frac{(\ln x_i-\mu)^2}{2\sigma^2}\right] - n\ln\sigma.$$

The MLE for μ, denoted $\hat{\mu}$, can be found by differentiating l with respect to μ and setting the result equal to zero:

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{n}\frac{\ln x_i - \mu}{\sigma^2} = 0,$$

$$\frac{1}{\sigma^2}\left[\sum_{i=1}^{n}(\ln x_i) - n\mu\right] = 0.$$

Multiplying both sides by σ^2 and dividing by n shows that $\hat{\mu}$ can be found by calculating the mean of the natural logarithms of all of the observations in the dataset:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i.$$

Similarly, the MLE for σ^2, denoted $\hat{\sigma}^2$, can be found by differentiating l with respect to σ and setting equal to zero. The function is differentiated with respect to σ rather than σ^2 for simplicity - finding the MLE for σ is equivalent to finding that of σ^2.

$$\frac{\partial l}{\partial \sigma} = \sum_{i=1}^{n}\left[\frac{(\ln x_i-\mu)^2}{\sigma^3}\right] - \frac{n}{\sigma} = 0,$$

$$\frac{n}{\sigma} = \sum_{i=1}^{n}\left[\frac{(\ln x_i-\mu)^2}{\sigma^3}\right].$$

Multiplying both sides by σ^3 and dividing by n shows that $\hat{\sigma}^2$ can be calculated by finding the variance of the natural logarithms of all of the observations in the dataset:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \hat{\mu}).$$

To test whether a log-normal distribution would be a good fit to naturally occurring species abundance data, and to estimate appropriate parameters for the distribution, the OTU (99%) abundance data from the meiofauna community 18S dataset outlined in Section 2.2 was used. The values of the maximum likelihood estimates of the parameters μ and σ^2 for each site in the dataset can be seen in Table 4.2.

In addition to this, log-normal distributions were also fitted to a set of 16S data. These data were taken from a study on bacterial communities in the human gut (111) in which 12 different samples were taken. The maximum likelihood estimates for the log-normal parameters are shown in Table 4.3.

Site	PWK1	PWK2	PWK3	LH1	LH2	LH3	EGR1	EGR2	EGR3	MEye1	MEye2	MEye3
μ	2.32	3.05	2.10	2.86	1.77	1.91	1.48	1.90	1.38	1.71	1.63	1.85
σ^2	2.74	2.68	2.51	4.67	4.23	4.13	2.07	3.07	2.14	3.36	3.01	3.13
Site	SkyStaf1	SkyStaf2	SkyStaf3	HW1	HW2	HW3	DBay1	DBay2	DBay3	VNM1	VNM2	VNM3
μ	2.39	2.54	2.09	3.42	2.59	2.80	2.84	2.39	3.24	1.22	1.55	1.59
σ^2	4.22	3.21	3.89	2.56	2.29	2.78	3.30	4.72	4.38	2.79	3.53	2.93
Site	Mera1	Mera2	Mera3	CapFer1	CapFer2	CapFer3	Seah1	Seah2	Seah3	Exe1	Exe2	Exe3
μ	1.58	1.28	1.34	1.54	1.65	1.67	1.26	1.61	1.99	2.09	1.86	1.78
σ^2	2.70	3.21	3.61	2.17	2.41	2.42	2.29	3.08	5.09	3.69	3.09	2.84
Site	Porthw1	Porthw2	Porthw3	Sheer1	Sheer2	Sheer3	PrLimpa1	PrLimpa2	PrLimpa3	Sada1	Sada2	Sada3
μ	1.39	1.37	1.61	1.88	1.86	1.93	1.04	1.12	1.36	2.03	1.25	1.84
σ^2	4.12	4.42	4.88	4.23	3.26	3.08	2.18	3.47	3.77	3.01	2.42	3.33
Site	stJean1	stJean2	stJean3	Newb1	Newb2	Newb3	FirthF1	FirthF2	FirthF3	Fraser1	Fraser2	Fraser3
μ	1.90	1.74	1.69	1.36	1.23	1.76	0.79	1.35	1.95	1.72	1.77	1.61
σ^2	4.85	4.26	4.57	2.58	2.70	3.51	2.03	2.89	3.02	4.27	4.29	3.96
Site	FreshW1	FreshW2	FreshW3	Silecr1	Silecr2	Silecr3	Gamb1	Gamb2	Gamb3	Min	Max	Mean
μ	1.49	1.81	1.94	1.29	1.45	1.59	1.93	2.33	2.20	0.79	3.42	1.82
σ^2	3.31	4.62	3.47	2.28	2.70	2.80	3.76	3.85	4.33	2.03	5.09	3.35

Table 4.2: Fitted log-normal parameters for all sites in the meiofauna community dataset. Parameters are maximum likelihood estimates.

Sample	1	2	3	4	5	6	7	8	9	10	11	12	Mean
μ	0.71	0.70	0.98	0.85	1.08	0.97	0.87	0.93	0.82	1.14	1.05	1.09	0.93
σ^2	1.10	1.02	1.54	1.51	1.66	1.50	1.14	1.20	1.23	1.76	1.64	1.78	1.42

Table 4.3: Fitted log-normal parameters for all samples in the gut bacteria community dataset. Parameters are maximum likelihood estimates.

The mean values for the fitted log-normal parameters for the 18S data were found to be $\mu = 1.82$ and $\sigma^2 = 3.35$. However, it was decided that it may be sensible to choose values towards the lower end of the range for performance reasons because lower values for μ and σ generally produce a smaller overall abundance which allows the simulation software to run faster. The means of the fitted parameters for the 16S data were $\mu = 0.93$ and $\sigma^2 = 1.42$ which were lower than those for the 18S data. This further reinforced the decision to use lower values for the parameters to generate the *in silico* data.

The log-normal distributions using the mean parameters fitted to both datasets are shown in Figure 4.3. This plot shows that log-normal distributions with lower parameter values generally produce a large amount of random variables with lower values, representing lower abundances. The distribution with higher values for μ and σ^2 is flatter, so a higher proportion of high value random variables, representing higher abundances, will be yielded relative to distributions with lower parameter values.

When assigning log-normal distributions to the *in silico* datasets, a log-normal random vari-

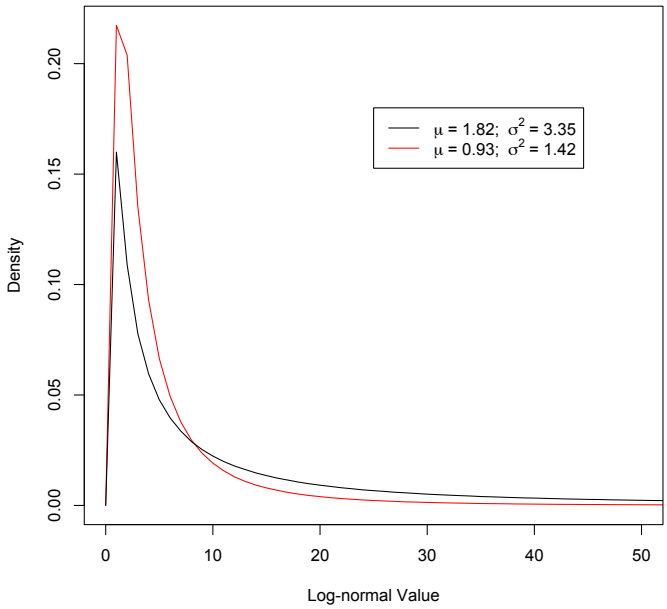

Figure 4.3: Probability density function for the log-normal distribution with two different sets of parameters. The black line shows the distribution using the mean parameters that were fitted to 18S data and the red line shows the distribution using the mean parameters that were fitted to 16S data. Note that the distribution is unbounded in the positive direction and, although the x axis of the plot is cut off at 50, it continues to infinity.

able with the decided upon parameters, X_i, was generated for each sequence, i. This random variable was then rounded up to the next integer and was assigned to sequence i as its abundance. This ensured that the most infrequently occurring sequences would have an abundance of 1 and the data would follow a discrete approximation to the log-normal distribution.

4.2.11 ROC Analysis

A *receiver operating characteristic* (ROC curve) is a plot of predictive data against known binary positive or negative values. The predictive data is typically a set of continuous variables (scores) representing the probability of a positive result for each data point. To initiate the analysis all data points are assumed to give a negative result (i.e. the threshold for a true result is lower than the lowest score in the set). This threshold is gradually increased and, as positive results are found, the proportion of true positives (the fraction of positive values that are identified as positive) is plotted against the proportion of false positives (the proportion of negative values that are identified as positive). This results in a monotonically increasing plot starting at (0,0) and ending at (1,1) and the area under this ROC curve (AUROC) assesses how good the predictive data performs. High AUROC values correspond to good predictions.

ROC analysis can be used to assess the performance of chimera detection software with a slight difference to the standard method in that the acceptance threshold is decreased instead of increased. For each sequence in a dataset, UCHIME outputs a probability that it is chimeric and, because the true nature of the simulated sequences is already known, a ROC analysis can be performed by gradually decreasing the UCHIME acceptance probability from high to low. Initially, all sequences will be considered good but as the threshold is lowered, sequences with UCHIME scores higher than this threshold will be flagged as chimeras. True positives are then plotted against false positives as normal.

4.3 Results

Group	Distribution	No. Sequences	Sample Size	Noise Added	No. Simulations
A1	$\ln N(0.79, 2.03)$	Variable	30000	None	30
B1	$\ln N(0.79, 2.03)$	7870	Variable	None	5
C1	Variable	7870	30000	None	25
D1	$\ln N(0.79, 2.03)$	7870	15000	Yes	5

Table 4.4: Summary of all *in silico* datasets - Greengenes.

Group	Distribution	No. Sequences	Sample Size	Noise Added	No. Simulations
A2	$\ln N(0.79, 2.03)$	Variable	30000	None	30
B2	$\ln N(0.79, 2.03)$	8000	Variable	None	5
C2	Variable	8000	30000	None	25
D2	$\ln N(0.79, 2.03)$	8000	15000	Yes	5

Table 4.5: Summary of all *in silico* datasets - Silva.

4.3.1 Summary of Datasets Analysed

A variety of different datasets were assembled using the methods described in Section 4.2. These datasets were designed in such a way that the effects of various attributes of microbial communities could be examined independently in order to form conclusions about the main influences on chimera formation and also on inferred community makeup. The different datasets which were created are summarised in Tables 4.4 and 4.5 and are described in this section.

Group A - Variation of Initial Species Richness

The effect of the richness of a sample was investigated by varying the initial number of species represented in the input data. Sequences were randomly sampled from both of the

in silico datasets to generate subsets containing 500 sequences, 1000 sequences, 2000 sequences, 4000 sequences and 6000 sequences in addition to the full datasets (7870 sequences for the Greengenes *in silico* data set and 8000 for the Silva *in silico* dataset). Each dataset was assigned an abundance distribution made up from $\ln N(0.79, 2.03)$ random variables, where the parameters were taken from the log-normal distribution fitted to the abundance data from Firth of Forth site 1. This process was repeated 5 times, producing 60 unique datasets - 30 made up of sequences from the Greengenes database and 30 made up of sequences from the Silva database.

25 rounds of PCR were simulated 5 times for each dataset. After PCR was simulated, 30000 sequences were sampled from the full pool of sequences to represent those that were detected during sequencing.

Note that, because all of the subsets were assigned the same abundance distribution, reducing the initial number of sequences in a dataset also reduces the initial overall abundance of that dataset by the same factor.

Group B - Variation of Sample Size

To investigate the effects that the number of reads yielded from sequencing has on the interpretation of the output data, subsamples of different sizes were drawn from simulated data. The full *in silico* datasets (7870 sequences for Greengenes and 8000 for Silva) were used with a $\ln N(0.79, 2.03)$ abundance distribution assigned. 25 rounds of PCR were simulated 5 times for both databases and for each simulation, the output sequences were subsampled with samples ranging from 1000 sequences up to 100000.

Group C - Variation of Log-normal Parameters

Different parameters for the log-normal distribution were selected to examine how the behaviour of the simulated data changed with these parameters. Initially, one of the parameters μ or σ was varied whilst keeping the other constant and then both parameters were varied simultaneously. For all simulations the full sized datasets were used (7870 sequences from Greengenes and 8000 sequences from Silva). Each simulation was repeated 5 times with different random variables, representing the abundances of the sequences, generated for each repetition. After each simulation, 30000 sequences were sampled from the output pool of sequences.

Varying μ

The estimated log-normal parameters for the abundance data at each site in the meiofauna community dataset are shown in Table 4.2. The mean value for σ^2 is 3.35 and it was decided to keep this value constant while varying the value of μ for a set of simulations. The parameter, μ, represents the mean of the natural logarithm of random variables drawn from a log-normal distribution. The actual mean of the log-normal distribution is given by $e^{(\mu+\sigma^2/2)}$ so it can be seen that increasing μ (with fixed σ) will also increase the mean of the distribution.

The values chosen were $\mu = 0$, 1, 2, 3 and 4 which span the range of estimated μ values for the meiofauna community dataset (min. $\mu = 0.79$ and max. $\mu = 3.42$).

Varying σ

The same method that was applied in Section 4.3.1 was repeated but, this time, the value of μ was fixed at 1.82 (the mean value of μ in Table 4.2) whilst σ was varied. The parameter σ represents the standard deviation of the logarithm of the random variables drawn from a log-normal distribution. The variance of these random variables is given by $(e^{\sigma^2}-1)e^{(2\mu+\sigma^2)}$. Therefore, keeping μ constant and increasing σ will increase the variance of the distribution.

The values σ = 0.5, 1.0, 1.5, 2.0, and 2.5 were chosen which covered the range of parameter values fitted to the meiofauna community dataset (min. σ=1.42, σ^2=2.03 and max. σ= 2.26, σ^2=5.09).

Varying both μ and σ

It can be shown that increasing either of the parameters σ and μ, whilst keeping the other constant, increases the overall abundance of the dataset, A, where X_i is the log-normal random variable used to assign the abundance of sequence i and n is the number of sequences in the dataset.

$$\mathrm{E}[X_i] = e^{(\mu+\sigma^2/2)}, i = 1 \ldots n; \tag{4.2}$$

$$\mathrm{Var}[X_i] = (e^{\sigma^2} - 1)e^{(2\mu+\sigma^2)}, i = 1 \ldots n; \tag{4.3}$$

$$A = \sum_{i=1}^{n} X_i$$

$$\Rightarrow \mathrm{E}[A] = n\mathrm{E}[X_i] = ne^{(\mu+\sigma^2/2)}. \tag{4.4}$$

Using the above equation for $n = 8000$ sequences and fixed $\sigma = 1.83$ (the mean value of σ in the meiofauna community dataset), the expected value of A is 42687 for $\mu = 0$ and 2330625 for $\mu = 4$. Similarly, if $\mu = 1.82$ (the mean value of μ in the meiofauna community dataset) is fixed then the expected value of A is 55949 for $\sigma = 0.5$ and 1123767 for $\sigma = 2.5$. Clearly, increasing either of the parameters μ or σ has the effect of increasing the overall abundance of the dataset.

Section 4.3.1 describes the effect of changing the initial number of sequences in the dataset and, by association, the overall abundance. Although the results show no correlation between the initial number of sequences and the effectiveness of chimera removal, it is noted that a larger starting dataset will result in a smaller proportion of the good sequences being identified. To avoid confusing the effects of this with the effects of changing the log-normal parameters, it was desirable to keep the expected abundance of the dataset constant whilst varying μ and σ together.

If the mean values of μ and σ from the meiofauna community dataset are used in Equation 4.4 with $n = 8000$ then the expected abundance, $\mathrm{E}[A] = 263602$ is returned. Keeping this value constant, a relationship between μ and σ can be derived such that

$$\mu = \ln\left(\frac{\mathrm{E}[A]}{n}\right) - \frac{\sigma^2}{2},$$

$$\mu = \ln\left(\frac{263602}{8000}\right) - \frac{\sigma^2}{2},$$

$$\mu = 3.495 - \frac{\sigma^2}{2}.$$

Values of μ were calculated for σ=0.5, 1.0, 1.5, 2.0 and 2.5 to give 5 pairs of parameters that maintained the expected abundance at a constant value of 263602. Log-normal random variables (7870 for Greengenes and 8000 for Silva) were generated using each pair of parameters to assign abundance distributions to the full datasets. This was repeated 5 times, and 25 rounds of PCR were simulated for each of the 50 resultant datasets. 30000 reads were sampled after the PCR simulation for each dataset.

Values for μ and σ are shown in Table 4.6 along with the expectation and variance, calculated using Equations 4.2 and 4.3, of the random variables drawn from the resultant probability distributions. Note that $\mathrm{E}[X_i]$ remains constant - as should be the case if n and $\mathrm{E}[A]$ are constant - and $\mathrm{Var}(X_i)$ increases as σ increases. Also note that, although $n = 7870$ for half

of the datasets in this section, this number is close enough to 8000 for the resulting values of μ to be the same as those used in the datasets where $n = 8000$.

σ	μ	E$[X_i]$	Var(X_i)
0.5	3.37	32.95	308.4
1.0	2.99	32.95	1847.0
1.5	2.37	32.95	9215.3
2.0	1.49	32.95	57613.6
2.5	0.37	32.95	561331.9

Table 4.6: Values of σ, μ, E$[X_i]$ and Var(X_i) for the constant value E$[A] = 263602$. σ and μ are log-normal parameters, E$[X_i]$ is the expected value of an associated log-normal random variable, Var(X_i) is the variance of an associated log-normal random variable and E$[A] = n$E$[X_i]$ for $n = 8000$.

Group D - Adding Simulated Noise

The full *in silico* datasets (7870 sequences for Greengenes and 8000 for Silva) were used with a ln$N(0.79, 2.03)$ abundance distribution assigned. 25 rounds of PCR were simulated 5 times for both databases and for each simulation. 15000 sequences were sampled from the output data in order to allow the software used on these datasets to run quickly enough. It is shown later in this chapter that using a smaller sample size reduces the quality of the output data in terms of how accurately it can be analysed, however, it is possible to infer the results for larger sample sizes from the results obtained from these datasets. Each simulated dataset was treated in five different ways:

1. No noise added. This represents a sample where all noise, apart from the chimeras, has been removed 100% accurately.

2. PCR single base errors added. See Section 4.2.5.

3. Pyrosequencing noise added. See Section 4.2.6.

4. Both PCR single base errors and pyrosequencing noise added. This represents a sample which has not been treated for noise.

5. Noise added and then removed using Amplicon Noise. This represents a realistic sample that has been treated for noise but will still contain some noisy sequences.

Note that, because additional noise is simulated after the chimeras have been simulated, noisy sequences which have been generated using original chimeras are still considered to be chimeras for the purposes of this analysis - i.e. they will be considered true positives if detected by UCHIME.

4.3.2 Chimera Detection

ROC analysis, as described in Section 4.2.11, was used to analyse how effective chimera checking software was at detecting chimeras in each dataset. The fasta files of sequences output from the simulations were checked for chimeras using the UCHIME *de novo* approach, ROC curves were plotted and AUROC values were compared to perform the assessment.

Figures 4.4 and 4.6 show the ROC curves generated for datasets with varying initial species richness (Groups A1 and A2). The areas under the ROC curves (AUROC) are shown in Figures 4.5 and 4.7. There is some variance present in the values but the AUROC values calculated suggest that chimera detection becomes more difficult as the number of starting sequences increases with, in general, fewer chimeras classified correctly for richer datasets.

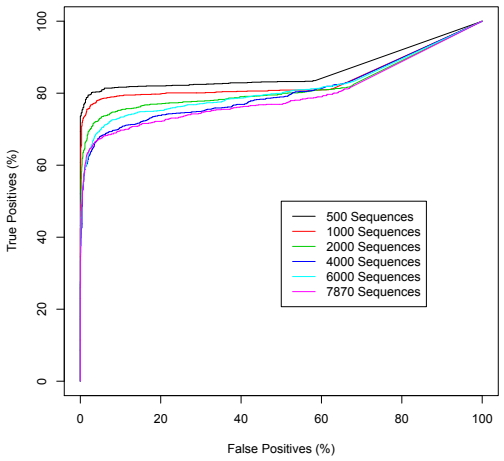

Figure 4.4: ROC curves to show effectiveness of chimera detection based on initial number of sequences in Group A1 datasets (see Table 4.4).

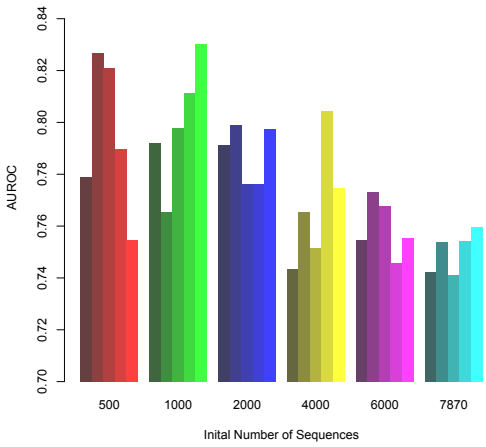

Figure 4.5: Areas under all ROC curves generated from Group A1 datasets (see Table 4.4). Each simulation was repeated 5 times.

Figures 4.8 , 4.9 , 4.10 and 4.11 show that UCHIME performed poorly when a small number of reads were sampled post-simulation but fared better when analysing larger samples. There is some variance in the data associated with which particular sequences were randomly sampled - some very obvious chimeras may have been selected in some samples but not in others - but the trend is, nevertheless, very noticeable.

For smaller samples, an increase in sample size improves chimera detection dramatically, but these improvements generally lessen in magnitude as larger samples are taken. Indeed, the difference in AUROC values between sample sizes of 1000 and 2000 is generally much more significant than the corresponding difference between sample sizes of 50000 and 100000. This effect can be seen in Figure 4.12 and suggests that the AUROC values are tending to-

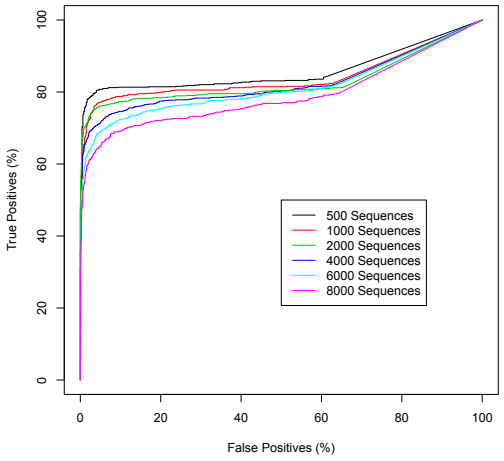

Figure 4.6: ROC curves to show effectiveness of chimera detection based on initial number of sequences in Group A2 datasets (see Table 4.5).

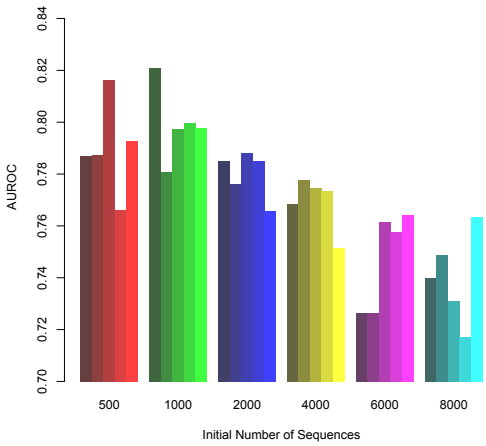

Figure 4.7: Areas under all ROC curves generated from Group A2 datasets (see Table 4.5). Each simulation was repeated 5 times.

wards a limit of around 0.8 for most of the datasets analysed in this section.

These results clearly show that, in practice, chimera removal will be more effective when using sequencing technologies that produce more reads - represented in these simulations by larger sample sizes.

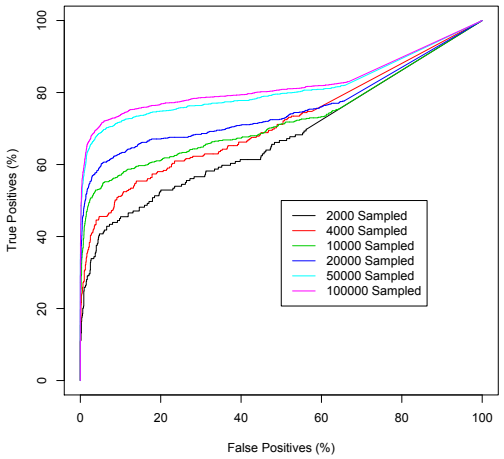

Figure 4.8: ROC curves to show effectiveness of chimera detection based on output sample size in Group B1 datasets (see Table 4.4).

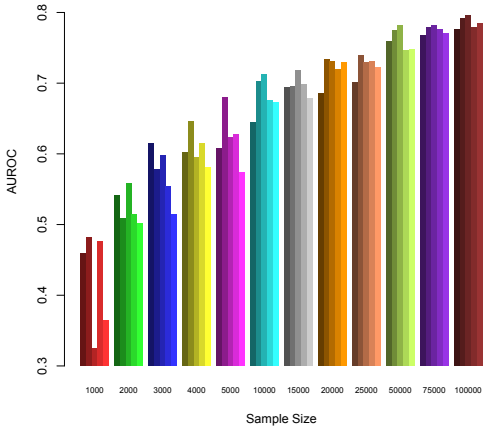

Figure 4.9: Areas under all ROC curves generated from Group B1 datasets (see Table 4.4). Each simulation was repeated 5 times.

It can be seen from Figures 4.13 and 4.14 that there is no obvious difference in the AUROC values calculated for various different values of μ, suggesting that this parameter does not affect the effectiveness of chimera detection using UCHIME.

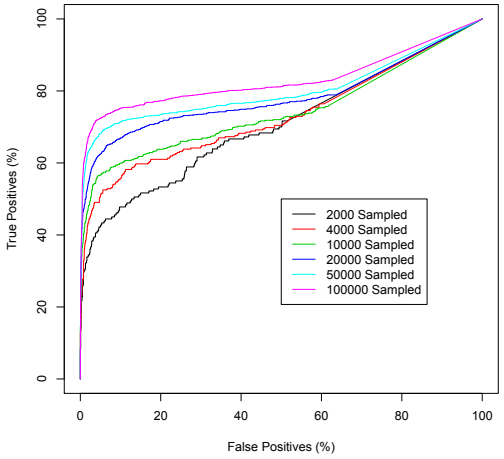

Figure 4.10: ROC curves to show effectiveness of chimera detection based on output sample size in Group B2 datasets (see Table 4.5).

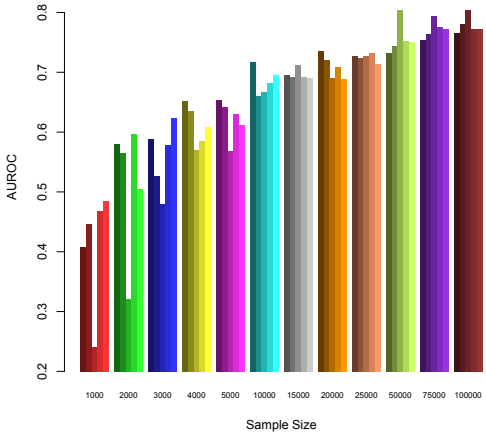

Figure 4.11: Areas under all ROC curves generated from Group B2 datasets (see Table 4.5). Each simulation was repeated 5 times.

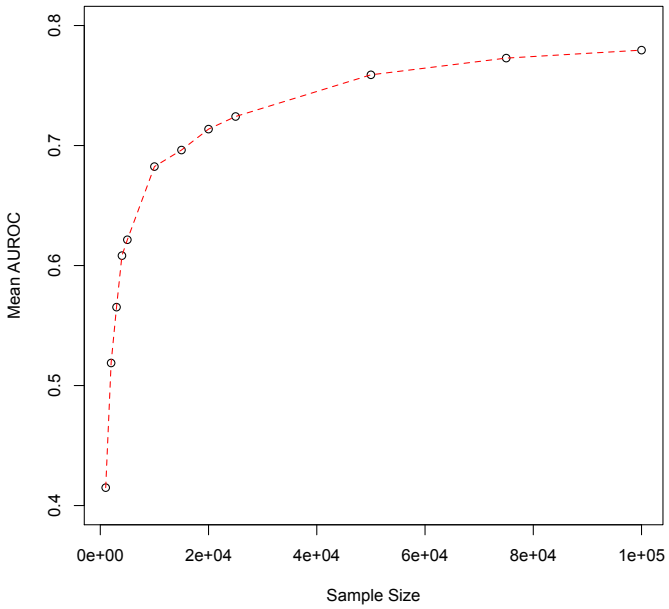

Figure 4.12: Mean AUROC values for all 12 sample sizes (1000 - 100000). Mean values are taken from a combination of all ROC curves, for that sample size, generated from both Group B1 and B2 datasets (see Tables 4.4 and 4.5).

The plots in Figures 4.15 and 4.17 demonstrate that chimera detection is more effective for datasets in which the sequence abundances are generated using higher values for σ and, therefore, have a higher variance. This can also be seen from the results in Figures 4.16 and 4.18 which also show that, generally, the effect of increasing σ lessens as σ increases. As in Section 4.3.1, where it was shown that the AUROC values tended towards an upper limit as the sample size increased, a similar effect can be seen as σ increases. Figure 4.19 shows the mean AUROC values seem to be tending towards a limit of approximately 0.8.

The AUROC values for data generated from log-normal distributions in which both parameters were varied together, shown in Figures 4.20 and 4.21, are similar to those calculated for the data where only σ was varied and μ remained constant (Figures 4.16 and 4.18). This suggests that, as expected, it is the parameter σ rather than μ or the total starting abundance that most influences the effectiveness of chimera removal software for log-normally distributed data.

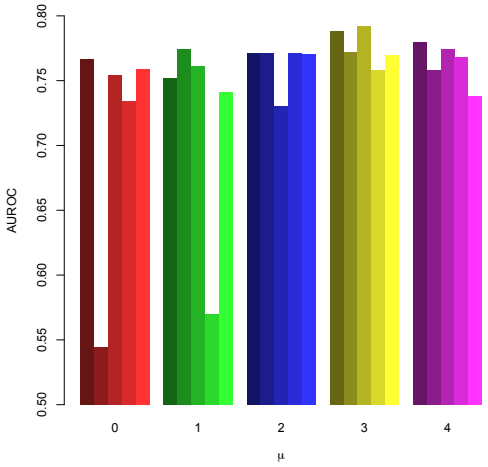

Figure 4.13: Areas under ROC curves generated from Group C1 datasets (see Table 4.4) with varying values for the log-normal parameter μ. Each simulation was repeated 5 times.

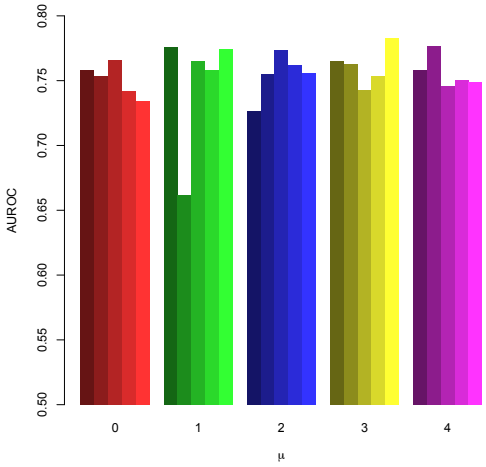

Figure 4.14: Areas under ROC curves generated from Group C2 datasets (see Table 4.5) with varying values for the log-normal parameter μ. Each simulation was repeated 5 times.

Figures 4.22 to 4.25 show the performance of UCHIME for different methods of simulating noise. It can be seen that in all cases, peak performance for chimera detection is reached when no additional noise is simulated, that is, for datasets in which non-chimeric noise has been 100% removed. This suggests that noisy data confuses the chimera checking algorithm which is logical because the chimeras and their respective parents are less likely to show similar characteristics with added noise.

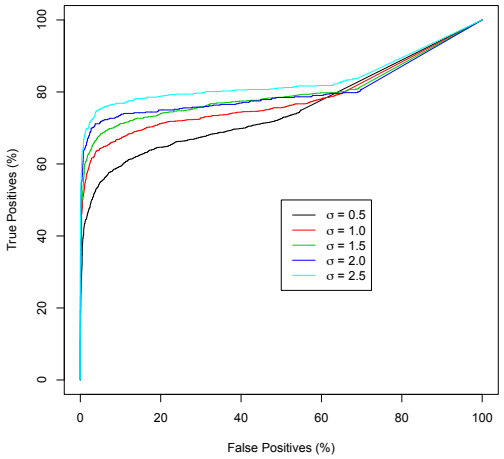

Figure 4.15: ROC curves to show effectiveness of chimera detection based on different values for the log-normal parameter σ in Group C1 datasets (see Table 4.4).

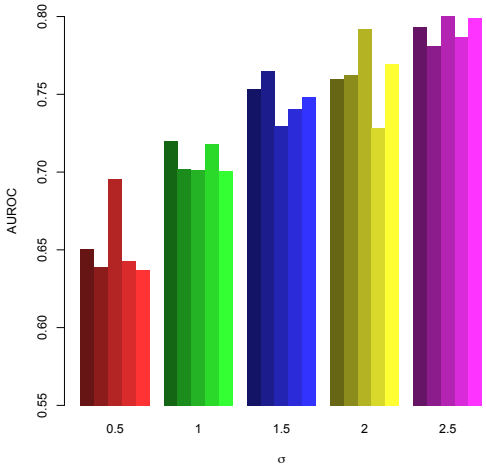

Figure 4.16: Areas under ROC curves generated from Group C1 datasets (see Table 4.4) with varying values for the log-normal parameter σ. Each simulation was repeated 5 times.

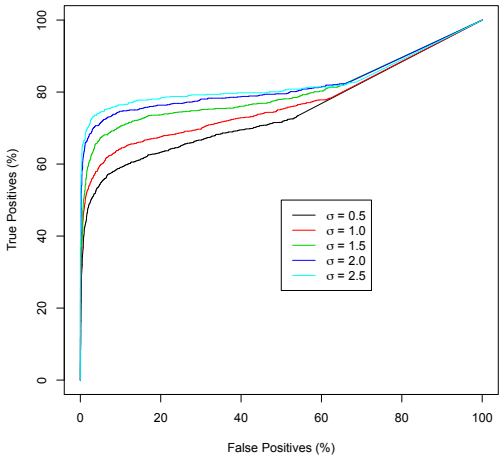

Figure 4.17: ROC curves to show effectiveness of chimera detection based on different values for the log-normal parameter σ in Group C2 datasets (see Table 4.5).

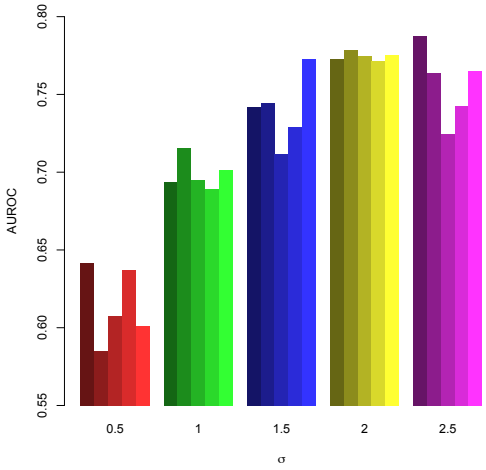

Figure 4.18: Areas under ROC curves generated from Group C2 datasets (see Table 4.5) with varying values for the log-normal parameter σ. Each simulation was repeated 5 times.

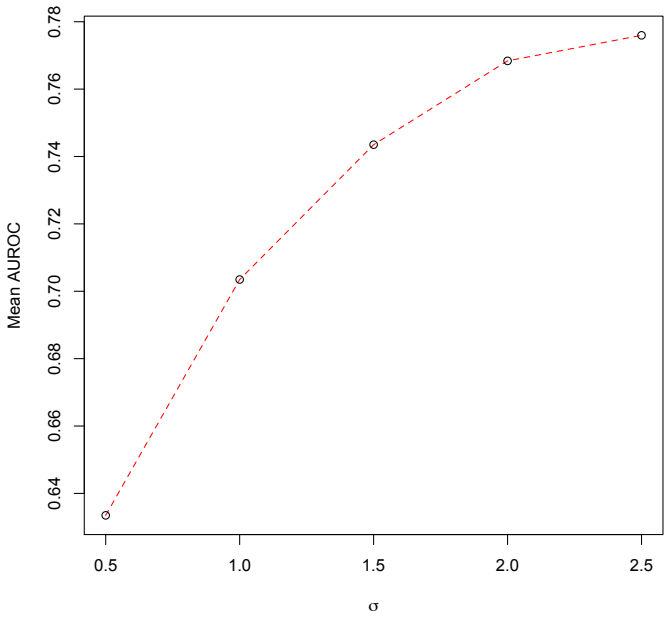

Figure 4.19: Mean AUROC values for all values of σ (0.5 - 2.5). Mean values are taken from a combination of all ROC curves, for that value of σ, generated from both C1 and C2 datasets (see Tables 4.4 and 4.5).

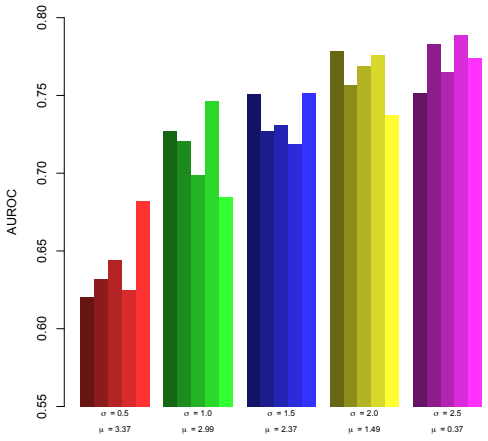

Figure 4.20: Areas under ROC curves generated from Group C1 datasets (see Table 4.4) with varying values for both of the log-normal parameters, μ and σ. Each simulation was repeated 5 times.

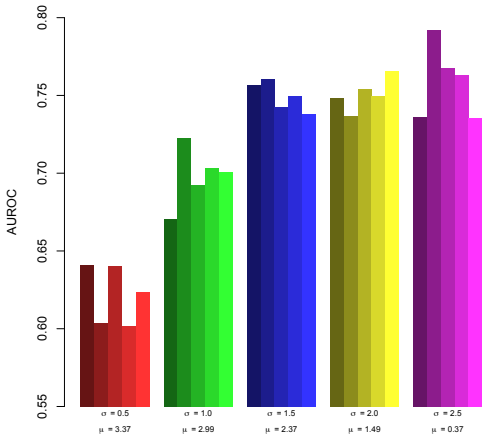

Figure 4.21: Areas under ROC curves generated from Group C2 datasets (see Table 4.5) with varying values for both of the log-normal parameters, μ and σ. Each simulation was repeated 5 times.

Simulated noise due to PCR single-base errors has much less of a negative effect on chimera detection than pyrosequencing noise. When pyrosequencing noise is added, either alone or in conjunction with other noise, AUROC values lower than 0.4 are calculated and, typically, more than 60% of chimeras are undetectable thus rendering the software too unreliable to be of any use. The poor results returned from datasets with simulated pyrosequencing noise relative to those with simulated PCR noise imply that UCHIME deals with transcription/translation errors better than insertion/deletion errors because the latter are created as part of pyrosequencing noise but not as part of PCR errors.

After the simulated noise has been removed it can be seen that UCHIME performance increases but not quite to the level of that observed on noise-free datasets, reaffirming that chimera detection is affected by the amount of other noise in the data. This demonstrates the importance of using an effective noise removal pipeline - poor performance in one area can have a knock-on effect and adversely impact performance in another area.

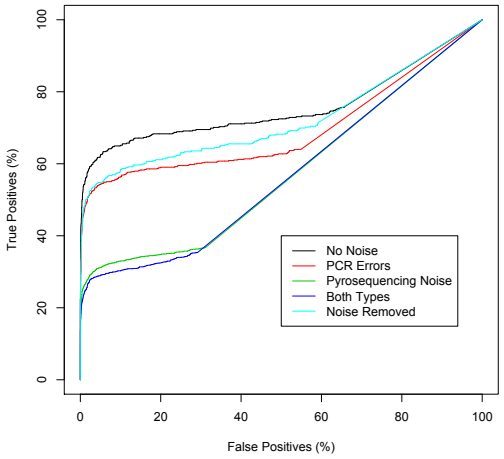

Figure 4.22: ROC curves to show effectiveness of chimera detection based on different methods of noise simulation in Group D1 datasets (see Table 4.4).

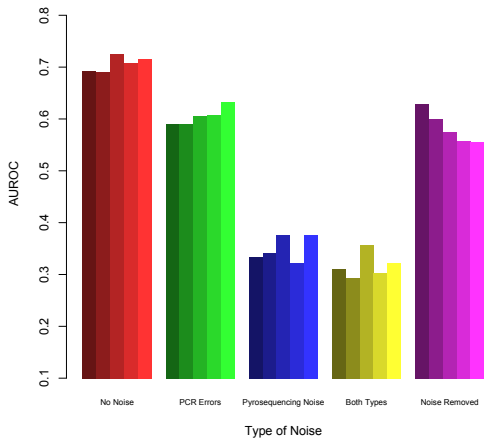

Figure 4.23: Areas under ROC curves generated from Group D1 datasets (see Table 4.4). Each simulation was repeated 5 times.

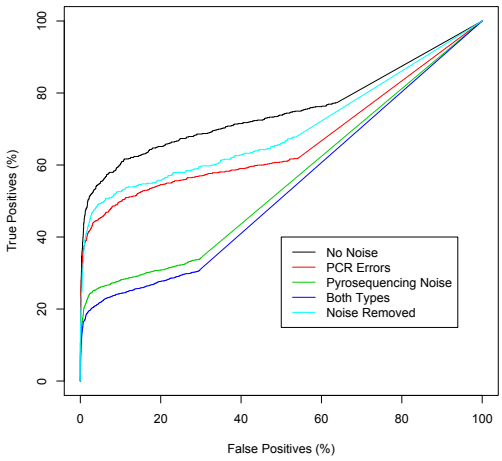

Figure 4.24: ROC curves to show effectiveness of chimera detection based on different methods of noise simulation in Group D2 datasets (see Table 4.5).

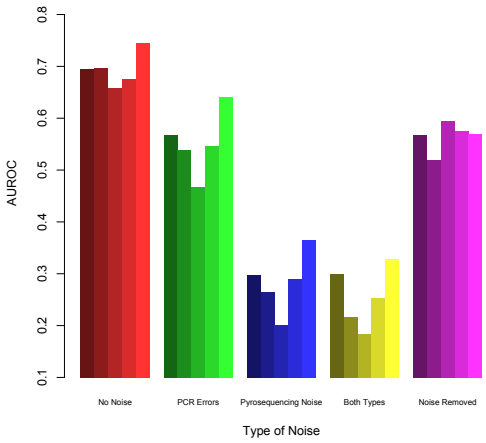

Figure 4.25: Areas under ROC curves generated from Group D2 datasets (see Table 4.5). Each simulation was repeated 5 times.

4.3.3 *De Novo* Chimera Detection Versus Reference-Based Chimera Detection

The UCHIME reference approach was analysed and compared to the results observed using the *de novo* approach that are shown in the previous section. For most of the analysis in this section, only the results using the ChimeraSlayer 16S reference database are shown. The effects of using different reference databases are discussed at the end of this section.

Figures 4.26 and 4.27 show the results on datasets with differing initial richnesses. It has already been shown that chimera detection is less reliable on richer datasets when the *de novo* approach is used. However, here it can be seen that varying the initial richness has little effect on UCHIME's performance when using the reference approach. The results suggest that, for richer data, the reference method may be the most sensible choice and that for data containing fewer OTUs (around 1000) the *de novo* approach will yield more accurate results.

The composition of the data seems to be more of a factor when using the UCHIME reference method. Using this method, datasets composed from the Silva database (Figure 4.26) produce worse results than those composed from the Greengenes database (Figure 4.27), suggesting that more of the sequences in the Greengenes datasets are present in the reference database. This means that although initial richness has an impact on which UCHIME approach works the best, it is also important to consider how comprehensively the chosen reference database is expected to cover the data to be analysed.

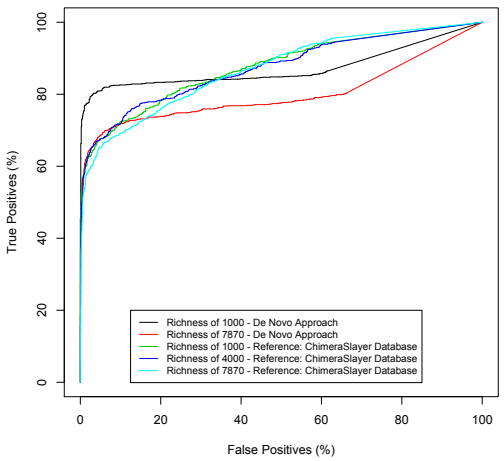

Figure 4.26: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group A1 datasets (see Table 4.4).

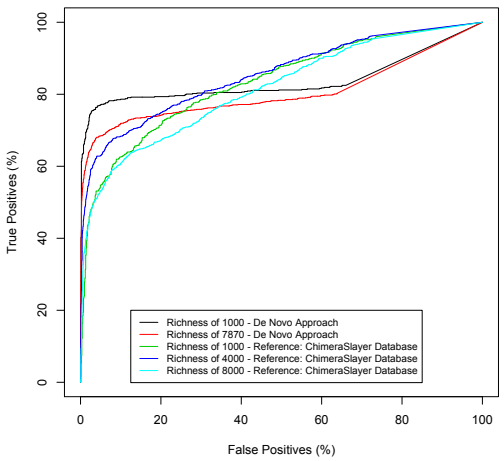

Figure 4.27: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group A2 datasets (see Table 4.5).

Figures 4.28 and 4.29 reveal a similar pattern when the sample size of the simulated output data is varied. As has been shown previously, larger sample sizes correlate with more accurate chimera detection when the UCHIME *de novo* approach is utilised. The reference-based approach, however, is shown to be a lot less dependent on the sample size and generally performs just as well for smaller samples. This suggests that for datasets with fewer reads, reference-based chimera checking should be employed and for datasets with more reads, the *de novo* approach is preferable.

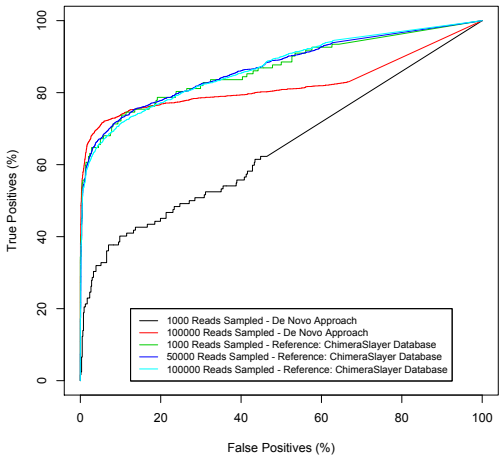

Figure 4.28: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group B1 datasets (see Table 4.4).

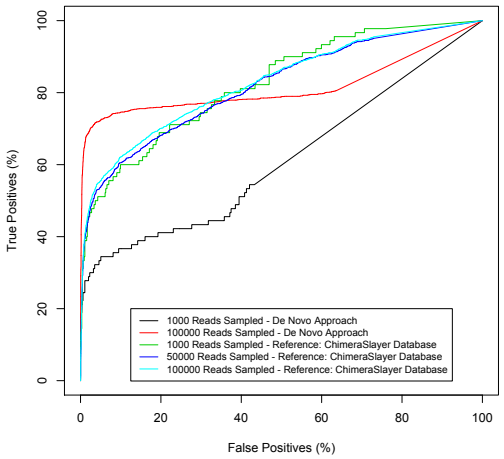

Figure 4.29: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group B2 datasets (see Table 4.5).

The same pattern continues in the results displayed in Figures 4.30 and 4.31 where the log-normal parameter σ and, consequently, the variance of the dataset has less of an effect on UCHIME's performance when the reference-based method is used in place of the *de novo* method. Therefore, a logical strategy would be to use the *de novo* approach for datasets with distributions conducive to good chimera detection, i.e. those with high σ and high variance, and to use the reference-based method when these conditions are not met. Data which are distributed log-normally with higher variance and higher σ will tend to contain a few outlying species with much higher abundances than the rest and, therefore, represent communities with lower evenness values.

Figures 4.32 and 4.33 demonstrate that the presence of noise adversely affects the quality of UCHIME results regardless of whether the *de novo* or reference-based approach is used. Therefore, it is not necessary to consider the level of noise present in the data and, instead, other factors should be taken into consideration when deciding which approach to use. Note that the data used to generate Figures 4.32 and 4.33 had a sample size of 15000 reads which

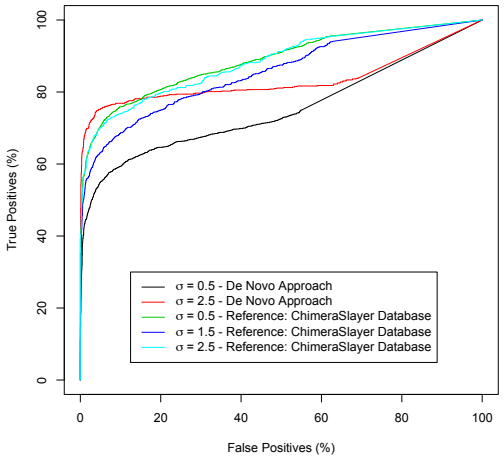

Figure 4.30: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection using different values for the log-normal parameter σ in Group C1 datasets (see Table 4.4).

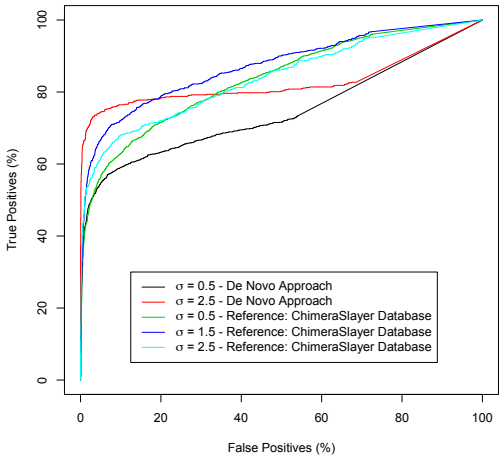

Figure 4.31: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection using different values for the log-normal parameter σ in Group C2 datasets (see Table 4.4).

causes the reference-based approach to produce better results in this instance.

Figures 4.34 to 4.37 compare the performance of the ChimeraSlayer reference dataset and the RDP classifier training dataset when used as reference datasets for the referenced-based approach in UCHIME. The RDP classifier dataset performs slightly better when used on simulated data generated from both Greengenes and Silva databases. This database contains almost twice as many sequences as the ChimeraSlayer database (10049 versus 5181sequences) so it is likely that it contained more reference sequences that matched with analysed data.

The reference based approach appears to work better on Greengenes data than it does on Silva data, again showing that some data may have poor representation in reference databases which could impact the performance of this method.

Note that the AUROC values are deceptive when comparing the two different UCHIME methods. For example, in Figures 4.36 and 4.37 the reference-based data return larger AUROC values but these can be misleading because much of the extra area is added when the acceptance threshold is very low - in this region the reference approach does identify more chimeras but also more false positives. Up to a more sensible threshold level the *de novo* approach is superior in this case.

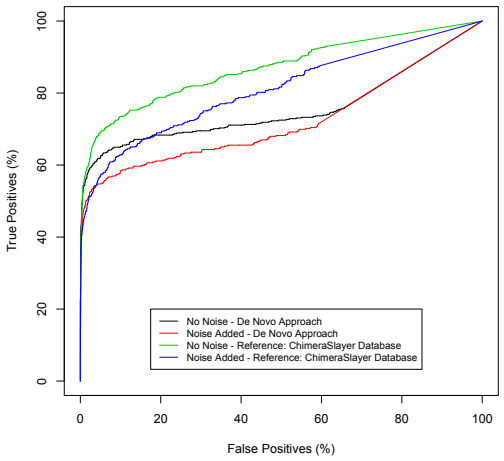

Figure 4.32: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group D1 datasets (see Table 4.4).

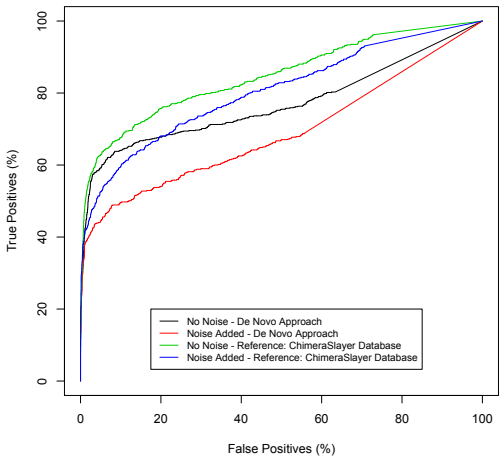

Figure 4.33: ROC curves comparing UCHIME *de novo* chimera detection with UCHIME reference-based chimera detection in Group D2 datasets (see Table 4.5).

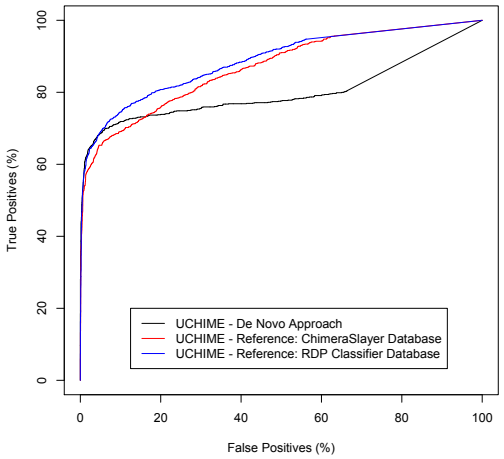

Figure 4.34: ROC curves to show effectiveness of chimera detection on *in silico* datasets generated from the Greengenes database. 30000 output sequences were sampled. The UCHIME *de novo* method was compared with the UCHIME reference method using two different reference databases.

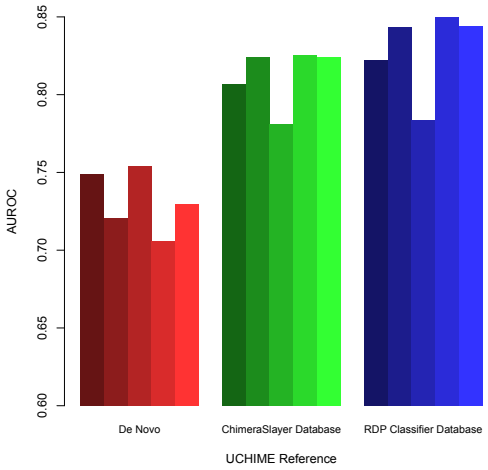

Figure 4.35: Areas under ROC curves generated from different methods of chimera detection on datasets generated from the Greengenes database. Each simulation was repeated 5 times.

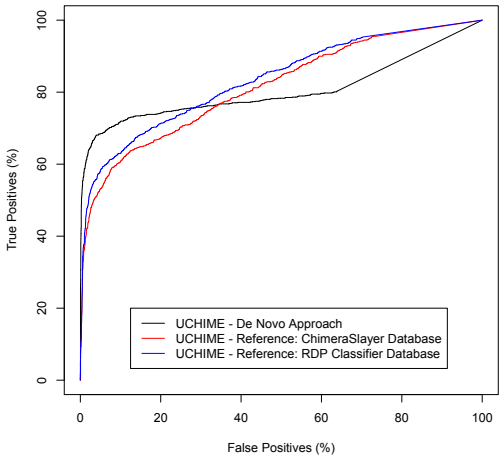

Figure 4.36: ROC curves to show effectiveness of chimera detection on *in silico* datasets generated from the Silva database. 30000 output sequences were sampled. The UCHIME *de novo* method was compared with the UCHIME reference method using two different reference databases.

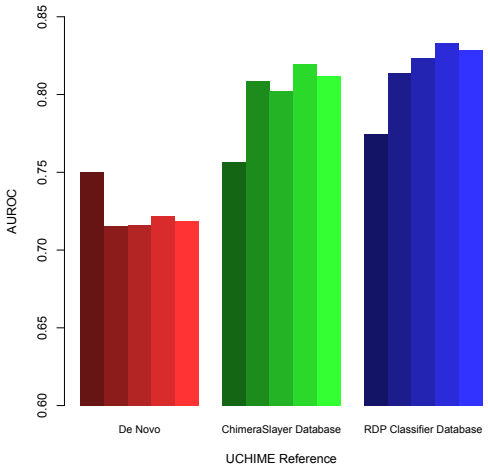

Figure 4.37: Areas under ROC curves generated from different methods of chimera detection on datasets generated from the Silva database. Each simulation was repeated 5 times.

4.3.4 UCHIME Versus Perseus - Chimera Detection

UCHIME was used for the majority of the testing of chimera detection software on the simulated datasets because, as explained in Section 4.2.8, UCHIME runs faster than Perseus and their performances have previously been found to be similar. In order to verify this, and to compare UCHIME against another method of chimera detection for simulated data of this type, Perseus was also tested on two opposing datasets in four areas of investigation - richness, sample size, variance of abundance distribution and the addition of noise. Datasets were identical to those in Table 4.4 with the exception that 15000 reads were sampled instead of 30000 to reduce the running time of Perseus.

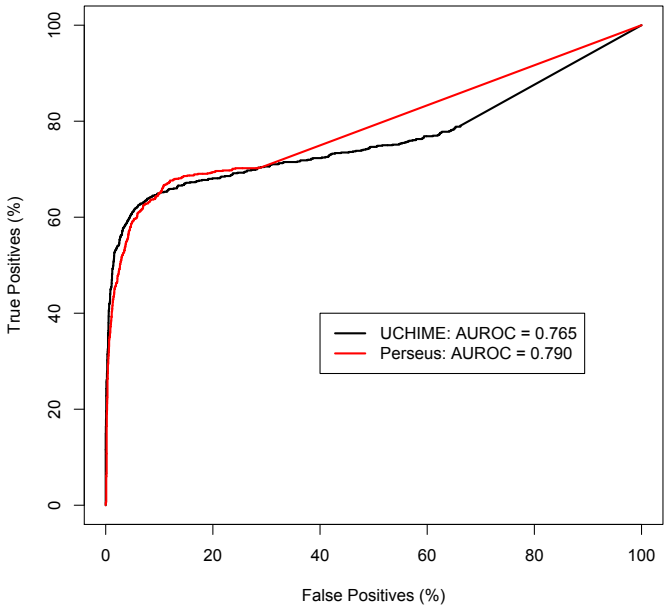

Figure 4.38: ROC curves to compare the effectiveness of chimera detection between UCHIME and Perseus on *in silico* noise-free datasets with relatively high richness, sample size and variance of abundance distribution. Datasets with initial richness of 7870, sample size of 15000 and abundance distribution with log-normal parameter $\sigma = 2.0$ were used. AUROC values are the mean of 5 replications of each dataset.

Figure 4.38 shows the results from "control" datasets with relatively high richness, abundance distribution variance, a high number of sampled read and no added noise. Subsequent analyses involve changes to these four variables and are compared to the results shown in this figure. Generally, Perseus performs comparably to UCHIME for datasets of this type with a slightly higher AUROC value. However it is to be noted that for a low false positive percentage, in the leftmost portion of the plot, UCHIME performs slightly better.

Comparing Figures 4.39 and 4.38 shows that Perseus performed better than UCHIME for

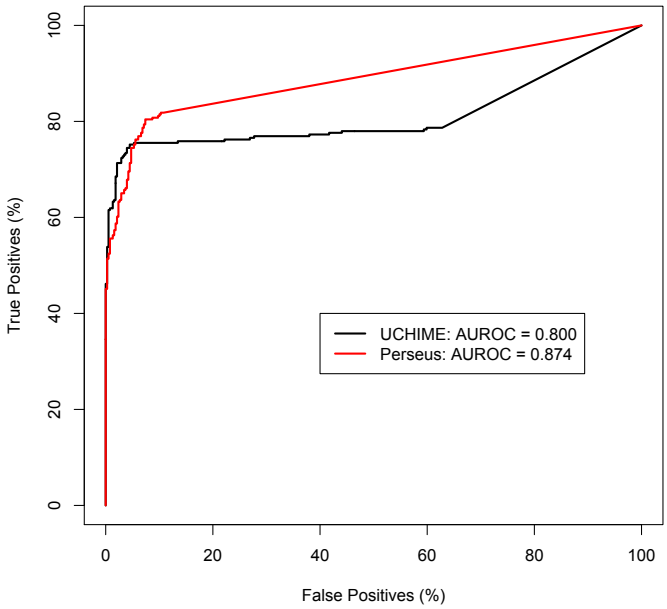

Figure 4.39: ROC curves to compare the effectiveness of chimera detection between UCHIME and Perseus on *in silico* datasets with relatively low richness. Datasets with initial richness of 500, sample size of 15000 and abundance distribution with log-normal parameter $\sigma = 2.0$ were used. AUROC values are the mean of 5 replications of each dataset.

datasets with 500 starting sequences and slightly worse than UCHIME when the full Greengenes based *in silico* dataset of 7870 sequences was used. As with UCHIME, Perseus performed better on datasets with lower richness.

When the sample size of the *in silico* datasets are varied, as shown in Figure 4.40, the results from Perseus again follow a similar pattern to those from UCHIME. In the case where 1000 reads were sampled from the simulated output data, Perseus performed very poorly, even compared to UCHIME, with the majority of chimeras undetected. With a higher sample size of 15000 (Figure 4.38), Perseus detected a larger number of chimeras and performed comparably to UCHIME.

In Figure 4.41 it can be seen that Perseus performed worse than UCHIME on datasets with low variance abundance distributions (log-normal parameter $\sigma = 0.5$) with the majority of chimeras undetected. Perseus performed much better on datasets with high variance abundance distributions (Figure 4.38) and the results were similar to those using UCHIME.

Figure 4.42 shows that, as with UCHIME, Perseus is less able to detect chimeras if noise is introduced to the data. The figure also suggests that Perseus deals with PCR and sequencing noise slightly worse than UCHIME because, even though Perseus returned a higher mean

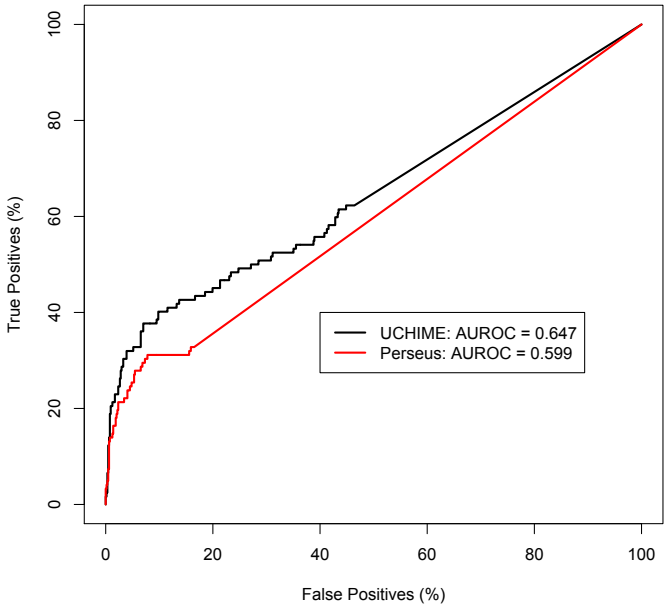

Figure 4.40: ROC curves to compare the effectiveness of chimera detection between UCHIME and Perseus on *in silico* datasets with relatively few sampled reads. Datasets with initial richness of 7870, a sample size of 1000 and abundance distribution with log-normal parameter $\sigma = 2.0$ were used. AUROC values are the mean of 5 replications of each dataset.

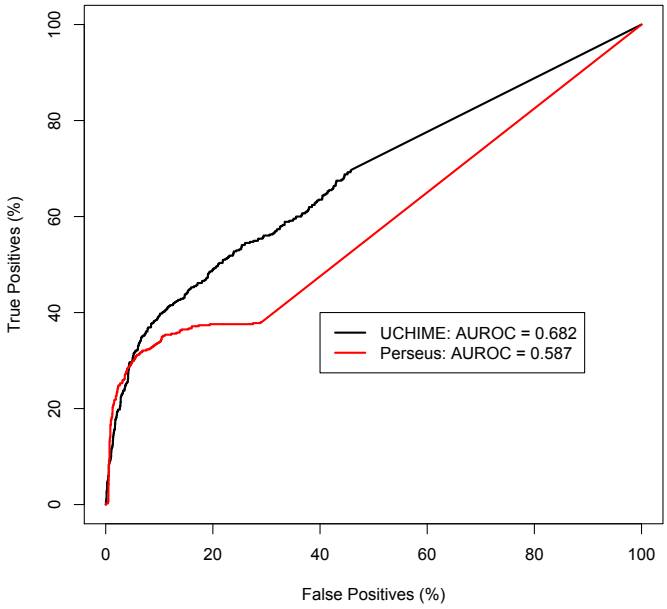

Figure 4.41: ROC curves to compare the effectiveness of chimera detection between UCHIME and Perseus on *in silico* datasets distributed log-normally with relatively low variance. Datasets with initial richness of 7870, a sample size of 15000 and abundance distribution with log-normal parameter $\sigma = 0.5$ were used. AUROC values are the mean of 5 replications of each dataset.

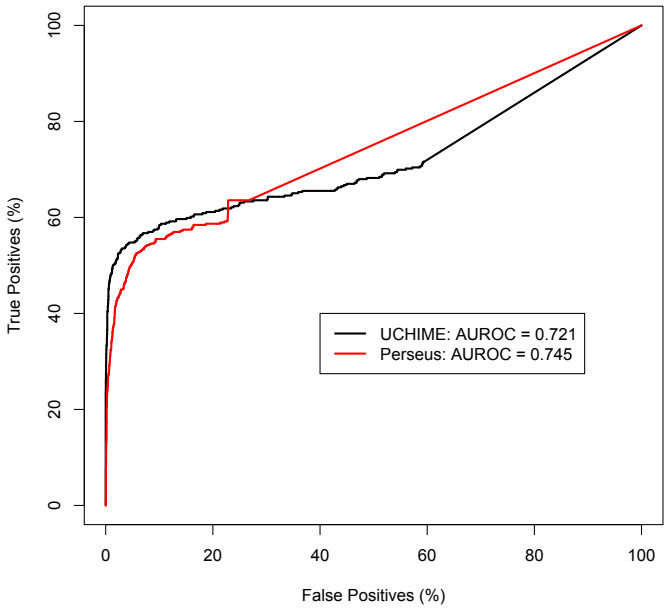

Figure 4.42: ROC curves to compare the effectiveness of chimera detection between UCHIME and Perseus on *in silico* datasets from which noise has been removed. The same datasets as those described in Figure 4.38 were used with the addition and subsequent removal of PCR noise and sequencing noise. AUROC values are the mean of 5 replications of each dataset.

AUROC value, UCHIME performed better for low false positive percentages in the region at the far left of the plot.

Overall, chimera detection with Perseus follows the same trends as with UCHIME. Chimera detection with Perseus is generally poorer when detection with UCHIME is poor and chimera detection is better when detection with UCHIME is better. In the case of datasets with low species richness, Perseus performs better than UCHIME; it is difficult to tell why this is the case but the Perseus algorithm may be better at selecting potential parent sequences from a smaller pool. In some cases UCHIME outperforms Perseus but in these cases the performance of UCHIME is generally very bad also. In other cases the two algorithms perform similarly. From the results it seems that it would be advisable to use Perseus on smaller datasets and UCHIME on larger datasets for reasons related to both performance and software running time.

4.3.5 Chimera Generation

The output files, including the various different sized samples of reads as well as the full output datasets recorded pre-sampling, from the simulations in this chapter were analysed in order to obtain information about how chimera generation is affected by the different variables chosen to generate the *in silico* data.

The data shown in Table 4.7 suggest that the initial species richness does not have any effect on the number of chimeras sequenced if the overall sampling size, in this case 30000, remains the same. This means that the overall percentage of chimeras will be higher in datasets with lower initial species richness and this is illustrated in Figure 4.43.

Another thing that is noticeable is that, as the number of initial sequences increases, a larger proportion of good sequences are missed during sequencing. This effect is portrayed in Figure 4.44 and can be simply explained by the fact that a fixed sample size will, clearly, give greater coverage on a smaller set. However, it is still important to note that, for large datasets, many organisms, particularly rarer ones, will be missed during sequencing.

Database Used	# Sequences (Initial)	# Chimeras	# Good Sequences	Good Sequences (% of Initial)
Greengenes	500	999.4	472.6	94.5
	1000	1025.8	923.6	92.4
	2000	828.8	1729.2	86.5
	4000	886.4	3094.6	77.4
	6000	747.2	4125.2	68.8
	7870	903.6	4973.4	63.2
Silva	500	895.8	469.8	94.0
	1000	946.0	932.4	93.2
	2000	869.0	1756.8	87.8
	4000	1033.4	3155.2	78.9
	6000	972.6	4294.0	71.6
	8000	827.0	5213.4	65.2

Table 4.7: Chimeras and good sequences sampled after PCR simulation on Group A1 and A2 datasets (see Tables 4.4 and 4.5). Data are mean values from the 5 replications of each dataset.

For datasets with varying sample size (Groups B1 and B2) it can be seen that as the sample size increases then the percentage of all good sequences included in the dataset increases (Figure 4.45) and the overall number of chimeras also increases (Figure 4.45). This was expected and is a consequence of larger sample sizes containing more sequences, however it demonstrates the lack of coverage associated with small sample sizes and the extra problem of more chimeras present in large sample sizes.

The percentage of all sequences that are chimeras generally rises as sample size increases, as shown in Figure 4.47. This is possibly caused by the dominance of high abundance good

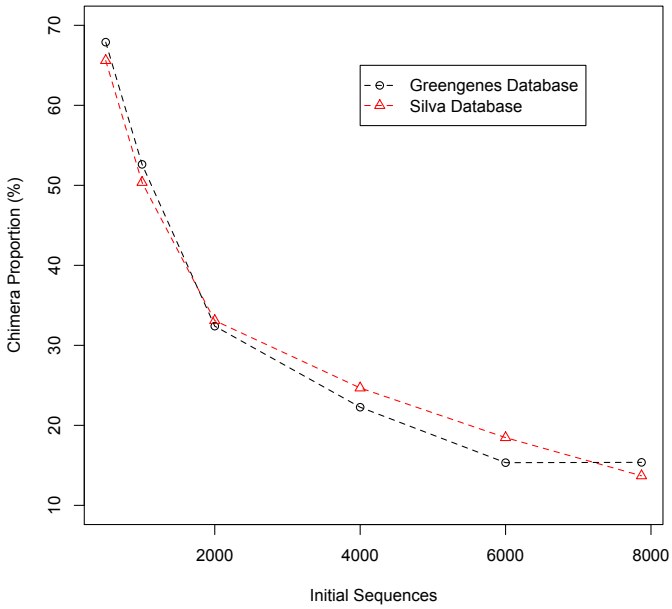

Figure 4.43: Plot to show the total percentage of all sampled sequences that were chimeras in Group A1 and A2 datasets (see Tables 4.4 and 4.5).

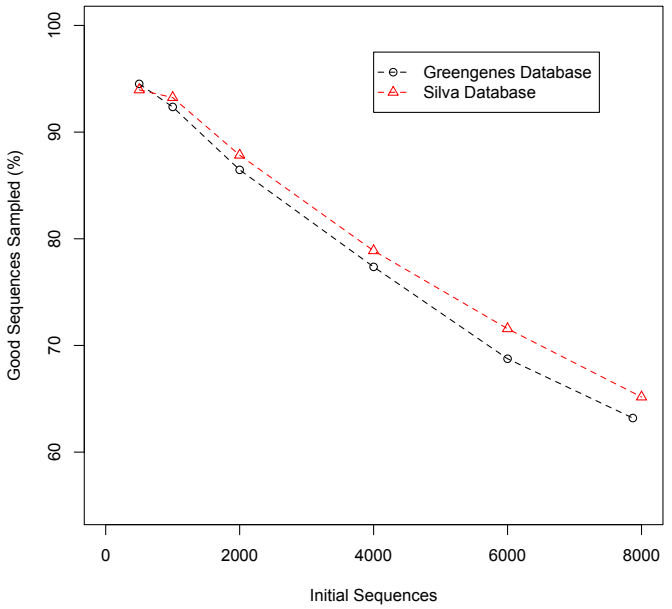

Figure 4.44: Plot of percentage of good sequences sampled against the initial number of sequences in Group A1 and A2 datasets (see Tables 4.4 and 4.5).

sequences in small samples. As sample size increases the rate of sampling of new good sequences is slower than the rate of sampling of new chimeras, increasing the chimera percentage in datasets with a large number of reads.

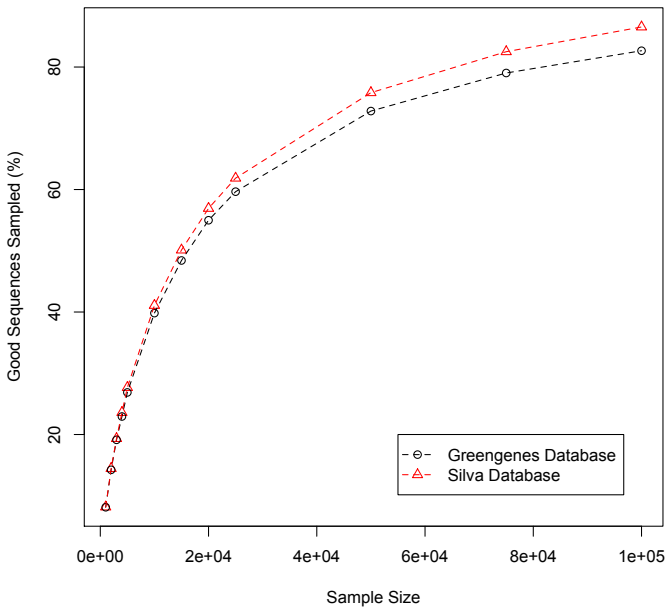

Figure 4.45: Plot of percentage of good sequences sampled against sample size in Group B1 and B2 datasets (see Tables 4.4 and 4.5).

The most noticeable result from the output data produced from datasets with varying log-normal parameters (Groups C1 and C2) was that the proportion of the good sequences sampled decreases as the value of σ increases (and μ decreases). This result can be seen in Figure 4.48 and there appears to be an almost linear relationship between the two variables.

The reason for this result is that, as σ increases, data drawn from the log-normal distribution tends to have a higher positive skew. This translates to the log-normal model for abundances by having relatively few sequences assigned with more of the total abundance. Therefore, for distributions with higher positive skew (higher σ), more reads will be sampled from these high abundance sequences and fewer sequences will be sampled overall.

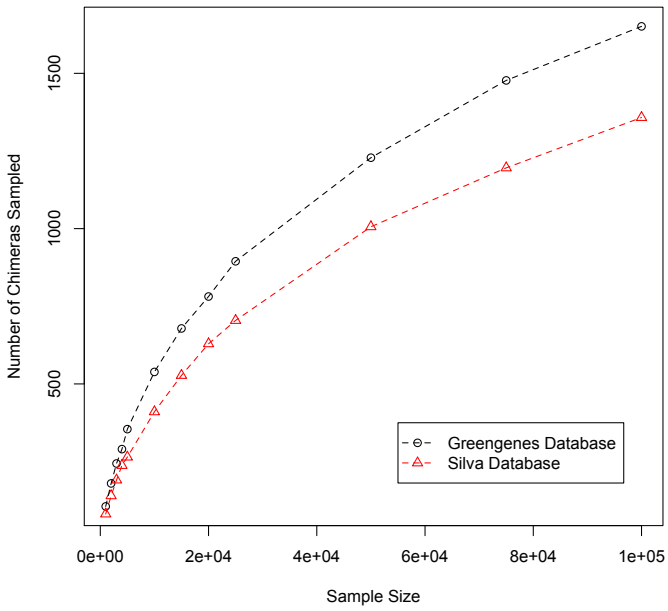

Figure 4.46: Plot of chimeras sampled against sample size in Group B1 and B2 datasets (see Tables 4.4 and 4.5).

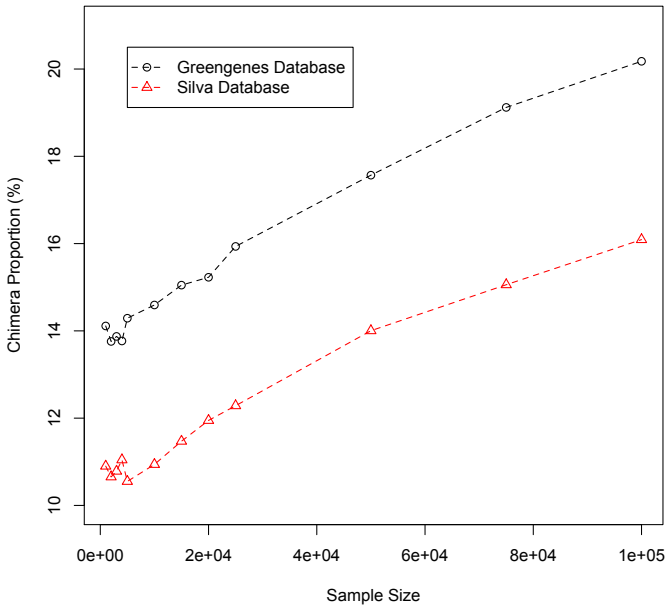

Figure 4.47: Plot to show the total percentage of all sampled sequences that were chimeras in Group B1 and B2 datasets (see Tables 4.4 and 4.5).

Both the overall number of chimeras formed in the full output dataset and the number of chimeras sampled from the output dataset also decrease as σ increases (Figures 4.49 and 4.50). Additionally, there is a decreasing trend between the proportion of sampled chimeras and σ, as can be seen in Figure 4.51. This is possibly because an evenly distributed dataset will increase the availability of a wider selection of potential parent sequences. A skewed dataset, conversely, will generally result in the repeated selection of the few high abundance sequences as parent sequences and this may lead to a higher frequency of independent formation of identical chimeras from the same parents.

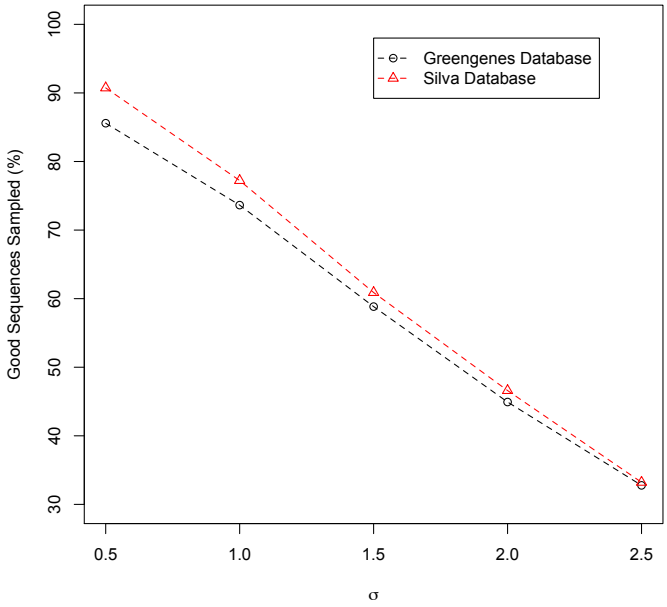

Figure 4.48: Plot of percentage of total good sequences sampled against the value of σ used to generate the abundance distribution in Group C1 and C2 datasets (see Tables 4.4 and 4.5). Values are the means of the 5 repeated simulations.

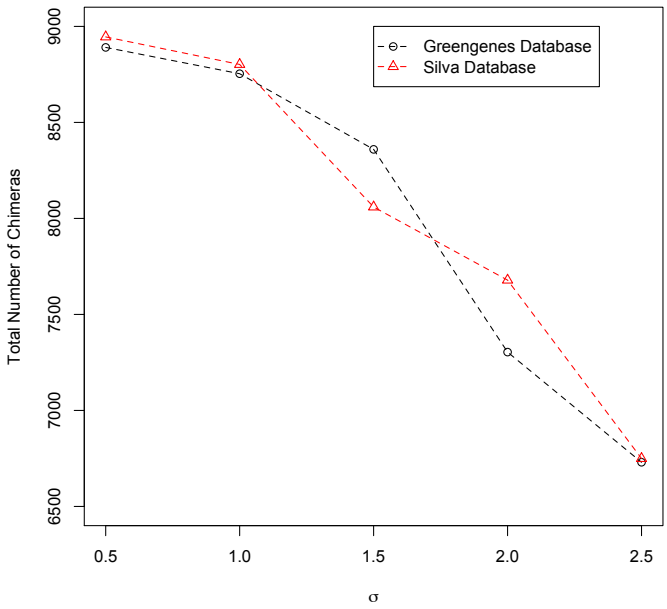

Figure 4.49: Plot of total chimeras generated against the value of σ used to generate the abundance distribution in Group C1 and C2 datasets (see Tables 4.4 and 4.5). Values are the means of the 5 repeated simulations.

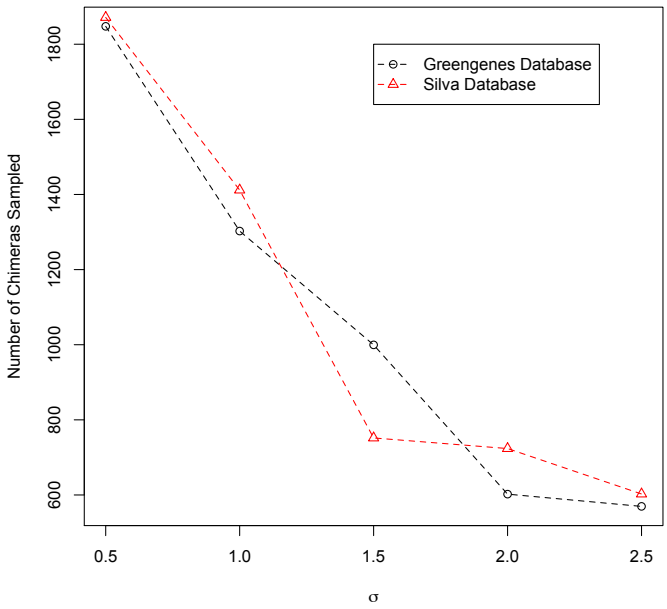

Figure 4.50: Plot of chimeras sampled against the value of σ used to generate the abundance distribution in Group C1 and C2 datasets (see Tables 4.4 and 4.5). Values are the means of the 5 repeated simulations.

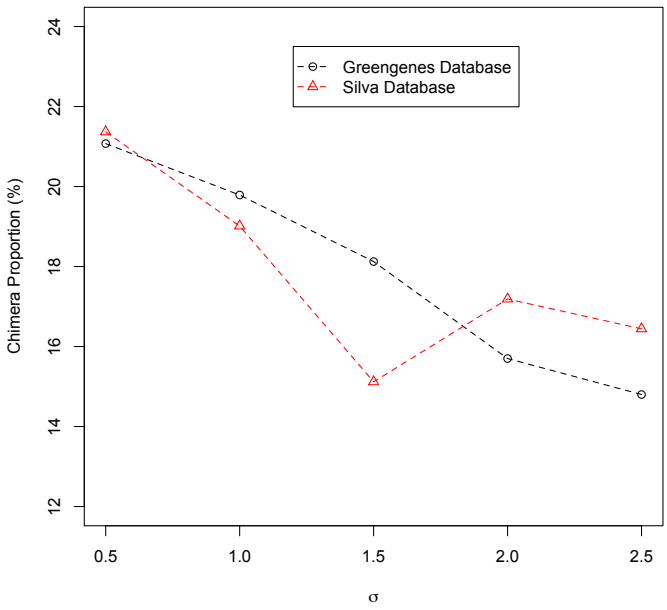

Figure 4.51: Plot of the total percentage of all sampled sequences that were chimeras against the value of σ used to generate the abundance distribution in Group C1 and C2 datasets (see Tables 4.4 and 4.5). Values are the means of the 5 repeated simulations.

4.3.6 Community Analysis

As the full input datasets and their associated output datasets could be analysed separately, it was possible to compare actual statistics relating to community composition of each dataset with those inferred from the output data after chimeras were removed using UCHIME in *de novo* mode. This allows conclusions to be drawn about what types of datasets yield results with the most accurately representative statistics and whether steps can be taken to account for any deficiencies inherent in datasets of certain types.

For all of the datasets that were analysed, the estimated richness (using the *Chao1* estimator), the Shannon diversity index and Pielou's evenness were calculated. Rarefaction analysis was performed where appropriate in order to determine how much of the full dataset remained hidden.

It can be seen in Figure 4.52 that for datasets with fewer species, richness is overestimated whereas for richer datasets, richness is underestimated when analysing the output data. However, particularly for datasets with around 4000 species or less, the *Chao1* estimator gives a good approximation to the true richness. Figures 4.53 and 4.54 suggest that both diversity and evenness are underestimated, with the error being greater in richer datasets for diversity and greater in less rich datasets for evenness.

The rarefaction curves in Figures 4.55 and 4.56 show that richer datasets generally have more of their richness hidden. Conversely, the flatter curves for datasets with fewer species show that most of the species are observed in the output generated from these datasets. In order to perform meaningful analysis of the vast majority of microbial communities, rich environmental samples are required, therefore the fact that these richer samples tend to produce less representative data is unfortunate.

As the output sample size increases (i.e. as the simulated number of reads increases), the estimated species richness increases, as can be seen in Figure 4.57. The gradient of the curve reduces as the sample sizes increase but it can be seen that the *Chao1* estimator greatly underestimates richness for smaller samples, particularly those with fewer than 10000 reads. A very similar pattern is observable in Figure 4.58 where the diversity of the input data is underestimated for smaller samples of the output data. Even the richness and diversity values that were calculated for the largest sample sizes were slight underestimates of the true measurements taken from the input data.

Figure 4.59 shows that calculated values of Pielou's evenness are higher, and generally closer to that of the input data, when smaller samples are taken from the output data.

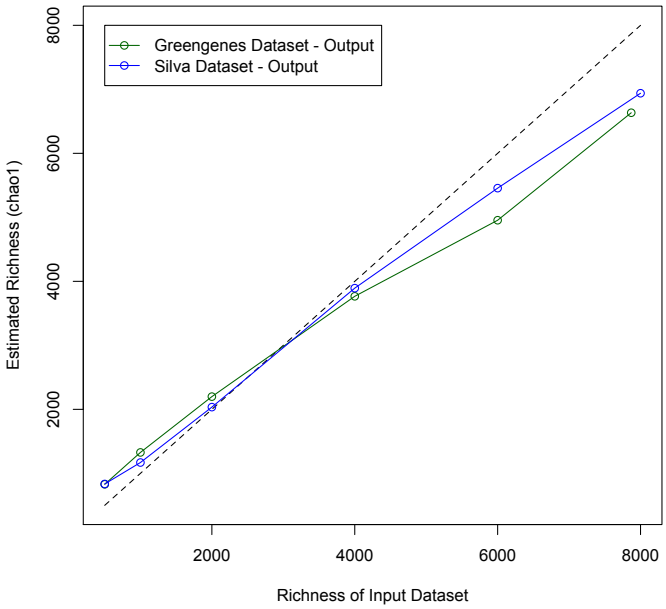

Figure 4.52: Group A1 and A2 datasets (see Tables 4.4 and 4.5): Initial species richness plotted against the estimated species richness, using *Chao1*, of the output data. Mean values for all simulations are shown.

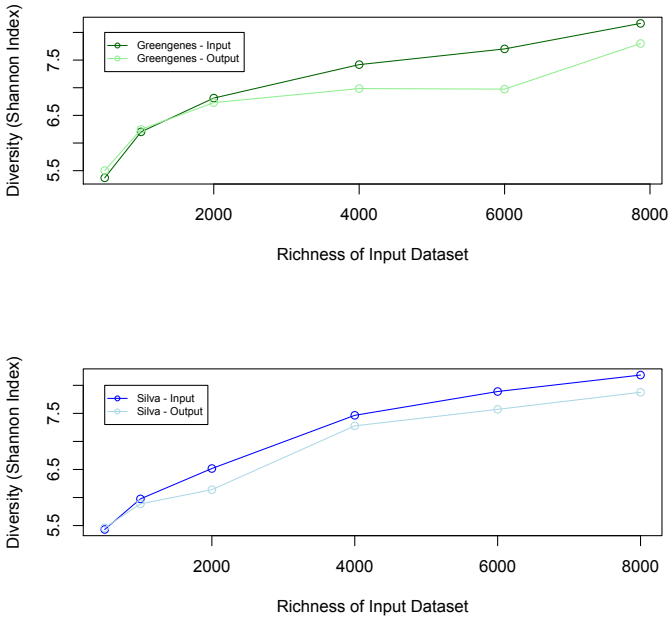

Figure 4.53: Group A1 and A2 datasets (see Tables 4.4 and 4.5): Initial species richness plotted against the Shannon diversity of both the input and output data. Mean values for all simulations are shown.

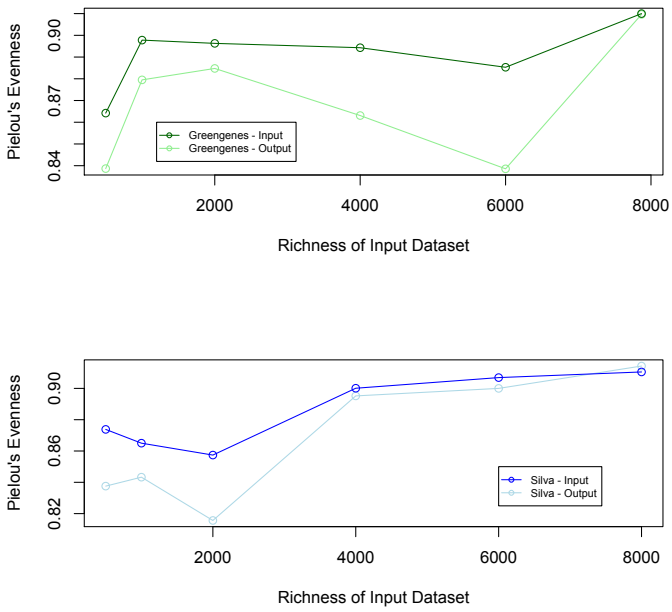

Figure 4.54: Group A1 and A2 datasets (see Tables 4.4 and 4.5): Initial species richness plotted against Pielou's evenness of both the input and output output data. Mean values for all simulations are shown.

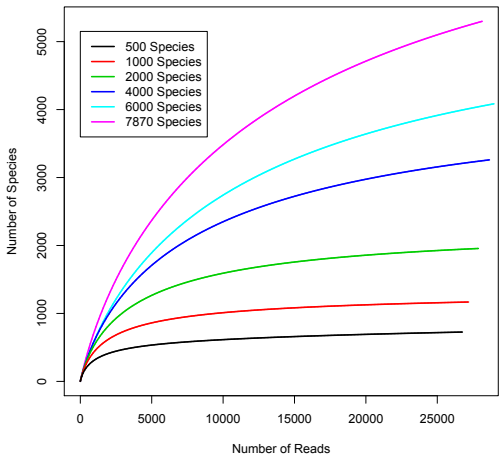

Figure 4.55: Group A1 datasets (see Table 4.4): Rarefaction curves using output data from simulations on datasets with varying initial species richness. The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

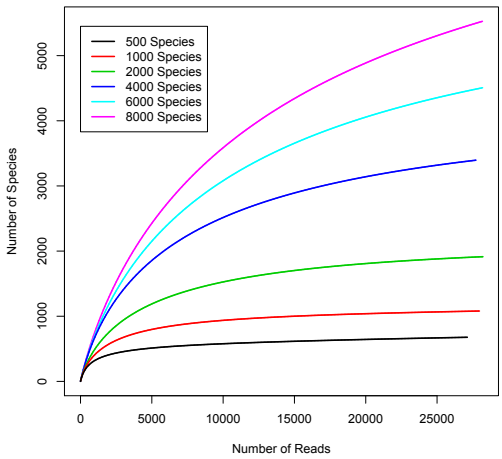

Figure 4.56: Group A2 datasets (see Table 4.4): Rarefaction curves using output data from simulations on datasets with varying initial species richness. The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

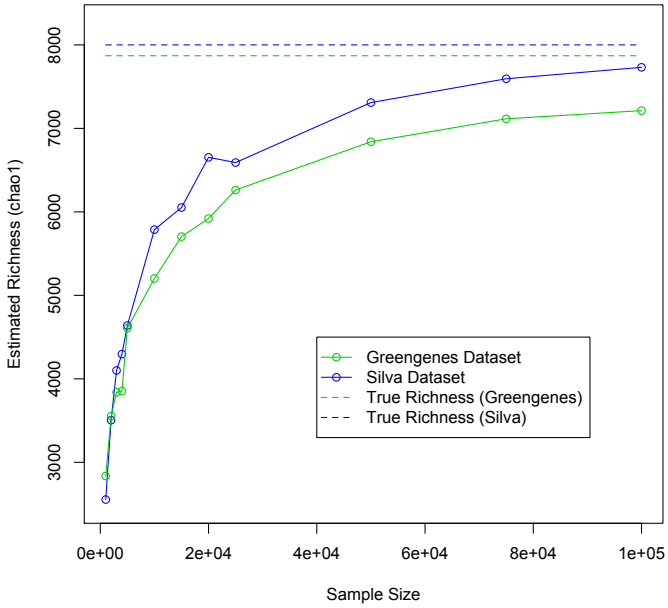

Figure 4.57: Group B1 and B2 datasets (see Tables 4.4 and 4.5): Output sample size plotted against the estimated species richness, using *Chao1*, of the output dataset. Mean values for all simulations are shown. The true richness values were 7870 for Group B1 datasets and 8000 for Group B2 datasets.

Whilst there is no evidence to suggest that the log-normal parameter μ has any effect on the observed composition of a dataset, the same cannot be said about the other log-normal parameter, σ. As has been discussed, a higher value of σ corresponds to higher variance which produces data with mostly species of low abundance but a few species of very high relative abundance which dominate the dataset. Lower values of σ produce flatter, more evenly distributed abundance data.

This misrepresentation of true richness can also be noticed in the rarefaction curves in Figures 4.63 and 4.64. The slopes of the curves featured here do not indicate that a greater amount of species richness is hidden in datasets with high σ than is hidden in datasets with low σ even though this is the case.

It can be seen in Figure 4.60 that higher variance datasets result in a big underestimation of the species richness, most likely caused by the small number of high abundance sequences dominating the sampled output, with fewer overall sequences selected. Flatter initial abundance distributions ($\sigma = 0.5$) tend to result in good estimates of the true richness.

As higher values of σ correspond to more uneven datasets, it follows that the input diversity will decrease as σ increases because the Shannon diversity is maximised when all species

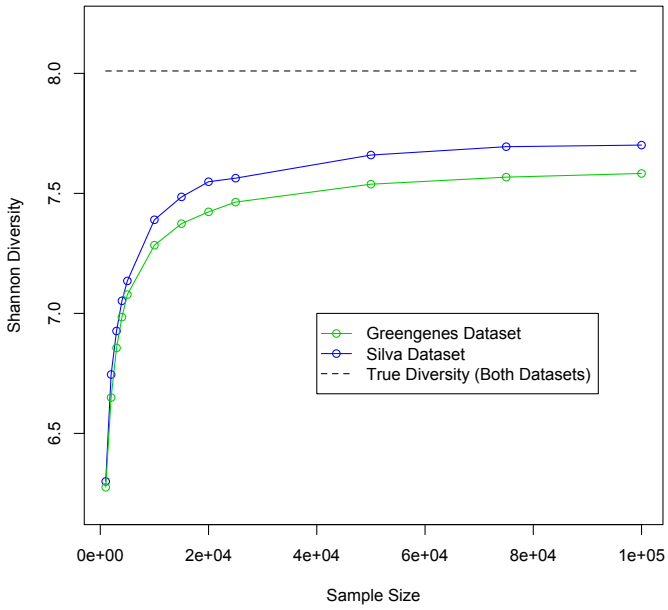

Figure 4.58: Group B1 and B2 datasets (see Tables 4.4 and 4.5): Output sample size plotted against the Shannon diversity of the output data. Mean values for all simulations are shown. The mean input (true) diversity values were 8.098 for Group B1 datasets and 8.101 for Group B2 datasets.

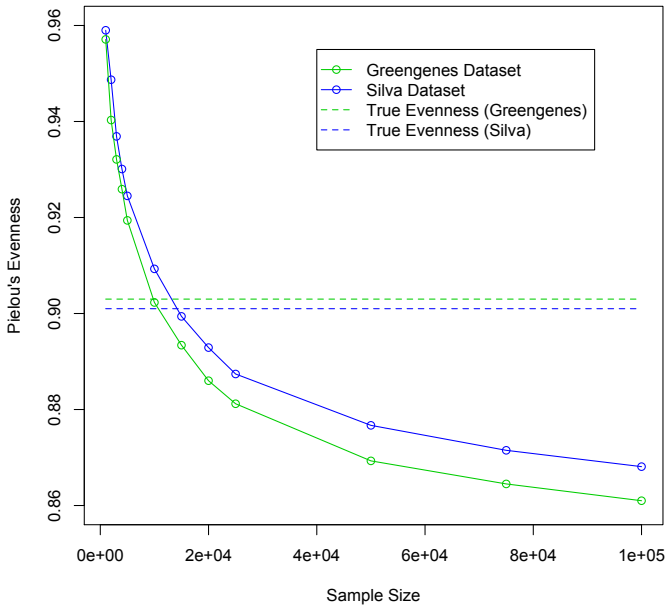

Figure 4.59: Group B1 and B2 datasets (see Tables 4.4 and 4.5): Output sample size plotted against Pielou's evenness of the output data. Mean values for all simulations are shown. The mean input (true) evenness values were 0.903 for Group B1 datasets and 0.901 for Group B2 datasets.

have the same abundance. This relationship can be observed in Figure 4.61 and is mirrored in the results for output diversity. However, the calculated values for output diversity underestimate the true diversity for all values of σ.

The unevenness of datasets with higher variance abundance distributions can be demonstrated by observing Pielou's evenness statistics (Figure 4.62). The evenness of the output data is generally a good representation of that of the input data but for higher σ values there is an overestimation. This could be due to the presence of low abundance chimeras in the output data, giving a greater number of reads with similar abundance.

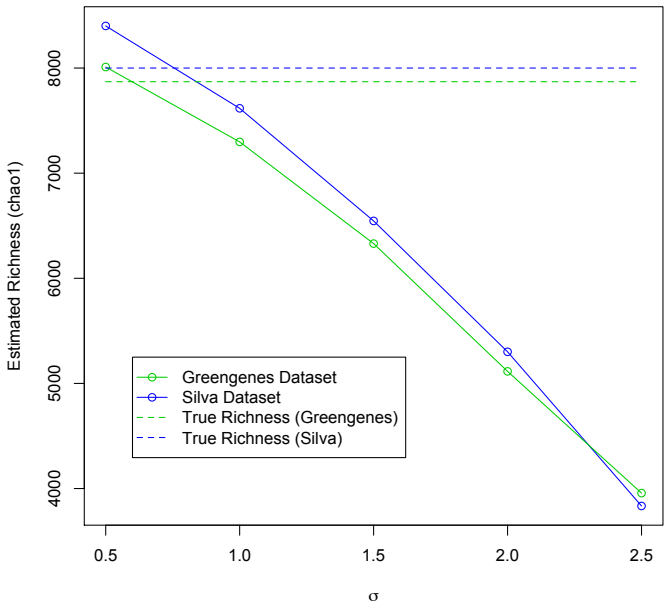

Figure 4.60: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Log-normal parameter σ (with constant $\mu = 1.82$) plotted against the estimated species richness, using *Chao1*, of the output data. Mean values for all simulations are shown. The true richness values were 7870 for Group C1 datasets and 8000 for Group C2 datasets.

Figures 4.65 to 4.67 show the results found when μ is varied as well as σ to keep the abundance of the input dataset constant. The results shown in these figures are very similar to those found in Figures 4.60 to 4.62 (where μ is kept constant while σ and the abundance are varied) which suggests that it is the parameter σ that is the overwhelming influence on the observed composition of log-normally distributed datasets and that any effects caused by μ and the initial abundance are negligible.

For analysis of datasets containing simulated noise (Group D1 and Group D2 datasets) it is apparent that the presence of pyrosequencing noise has a greater adverse effect on the ac-

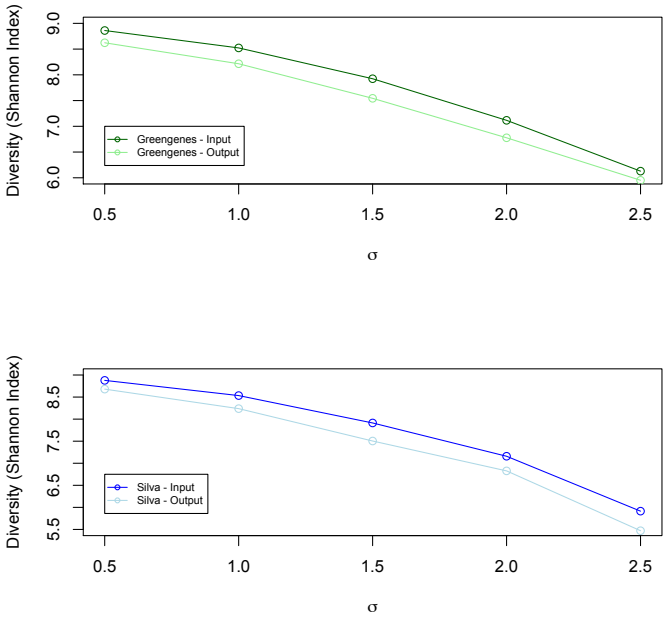

Figure 4.61: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Log-normal parameter σ (with constant $\mu = 1.82$) plotted against the Shannon diversity of both the input and output output data. Mean values for all simulations are shown.

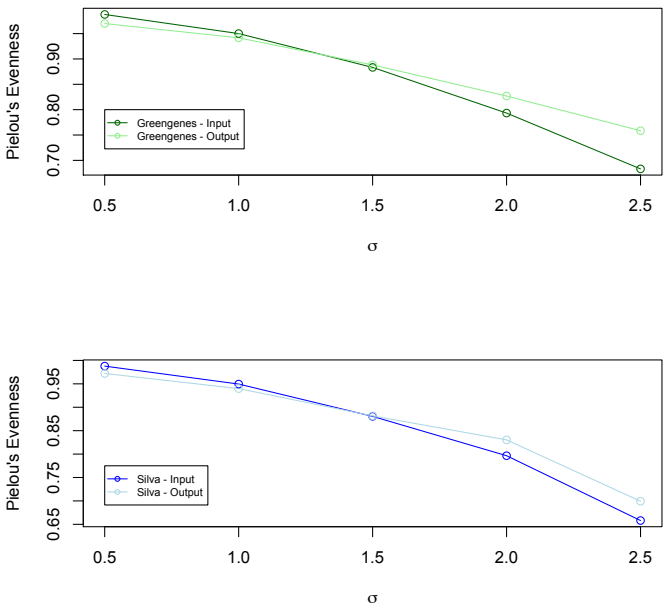

Figure 4.62: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Log-normal parameter σ (with constant $\mu = 1.82$) plotted against Pielou's evenness for both the input and output output data. Mean values for all simulations are shown.

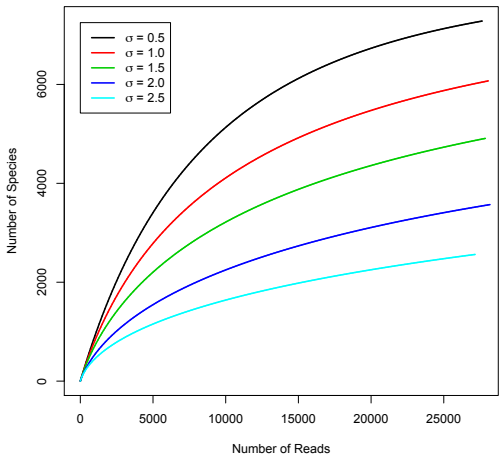

Figure 4.63: Group C1 datasets (see Table 4.4): Rarefaction curves using output data from simulations on datasets with varying log-normal parameter σ (with constant $\mu = 1.82$). The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

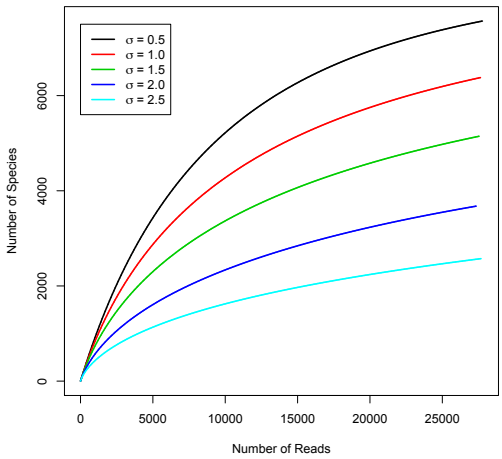

Figure 4.64: Group C2 datasets (see Table 4.5): Rarefaction curves using output data from simulations on datasets with varying log-normal parameter σ (with constant $\mu = 1.82$). The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

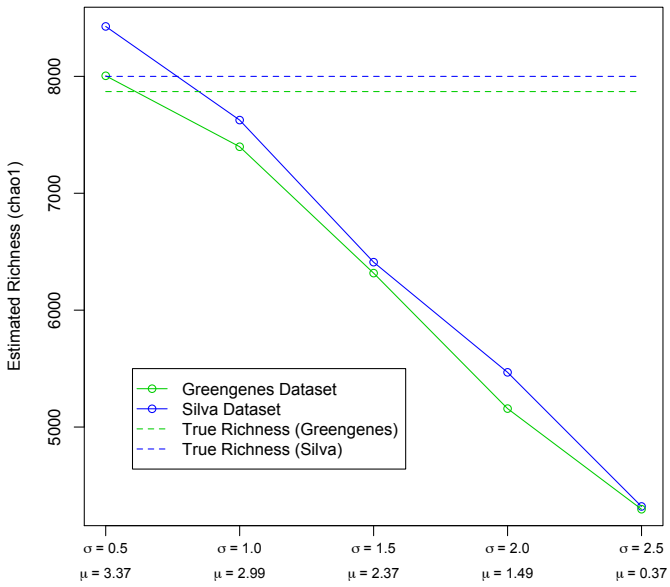

Figure 4.65: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Variable log-normal parameters σ and μ plotted against the estimated species richness, using *Chao1*, of the output dataset. Mean values for all simulations are shown. The true richness values were 7870 for Group C1 datasets and 8000 for Group C2 datasets.

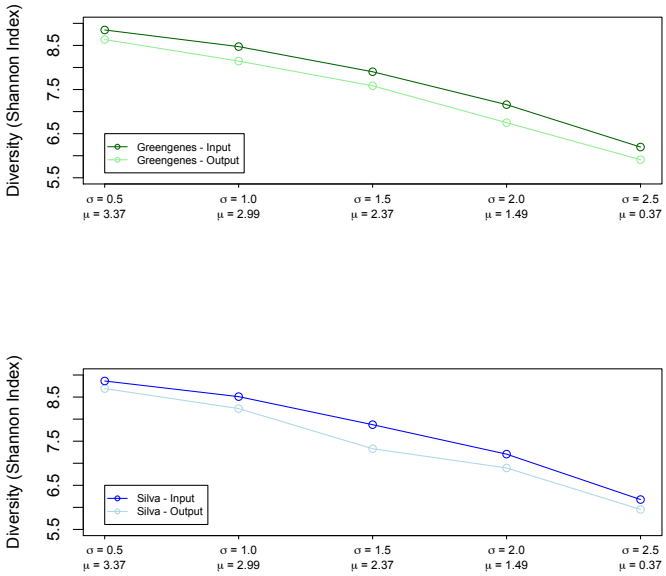

Figure 4.66: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Variable log-normal parameters σ and μ plotted against the Shannon diversity of both the input and output output data. Mean values for all simulations are shown.

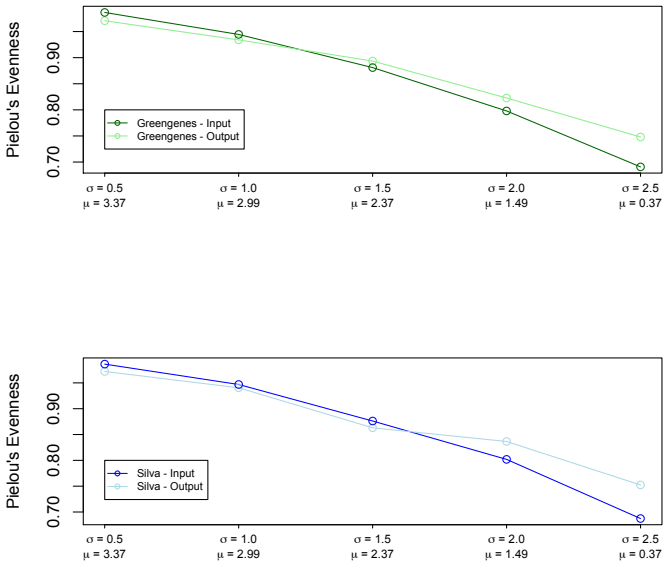

Figure 4.67: Group C1 and C2 datasets (see Tables 4.4 and 4.5): Variable log-normal parameters σ and μ plotted against Pielou's evenness for both the input and output output data. Mean values for all simulations are shown.

curacy of estimated community richness, diversity and evenness than the presence of PCR errors.

Figures 4.68 and 4.69 show that adding both types of noise overestimates community richness. This is, clearly, to be expected because the inclusion of noisy sequences artificially increases the number of species detected in a sample and, therefore, the observed richness. This will in turn affect the estimated richness using the *Chao1* estimator.

A comparison of data to which no noise has been added (representing perfect noise removal) with data from which the simulated noise has been removed reveals that the latter data produces better richness estimates. This is a by-product of the fact that richness estimates are generally underestimated for noise-free data, as shown earlier in this section, and, because of this, the residual noise is contributing to a better estimate. Ideally this would not be the case and the completely noise-free data would produce more accurate richness estimates.

As noisy datasets are of greater richness, it is likely that they will also have higher diversity and this is shown to be the case in Figures 4.70 and 4.71. An interesting observation from these results is that although the data with imperfectly removed noise yields higher richness estimates than those with perfectly removed noise, the opposite is the case for diversity measures - the data from which the simulated noise has been removed is less diverse.

These differences can be attributed to the evenness of the respective datasets (shown in Figures 4.72 and 4.73) because a community's diversity is a function of its richness and its evenness, thus the datasets from which simulated noise has been removed exhibit lower evenness. The relative unevenness of these data must be caused by the noise removal process, suggesting that some noisy sequences of high abundance may have been conserved in error and some good sequences of low abundance may have been discarded in error.

Figures 4.74 and 4.75 further show that the addition of pyrosequencing noise has a greater adverse effect than the addition of PCR errors. They also reiterate that almost half of a datasets richness will be missing when 15000 reads are samples. This is also the case for datasets from which simulated noise has been removed, with similar rarefaction curves for these datasets and those to which no noise was added.

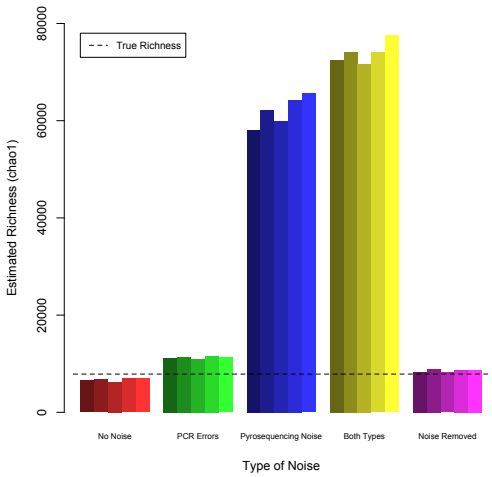

Figure 4.68: Estimated richness for all Group D1 output datasets (see Table 4.4). Each simulation was repeated 5 times.

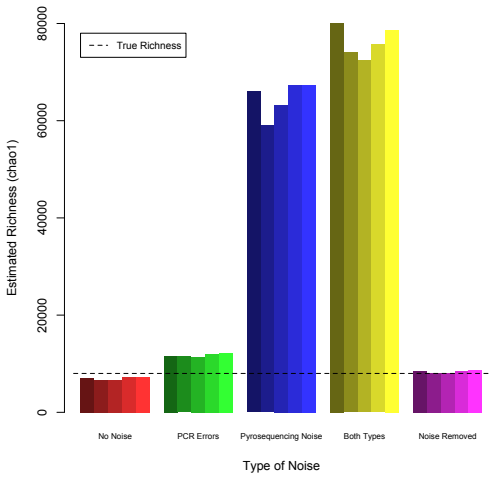

Figure 4.69: Estimated richness for all Group D2 output datasets (see Table 4.5). Each simulation was repeated 5 times.

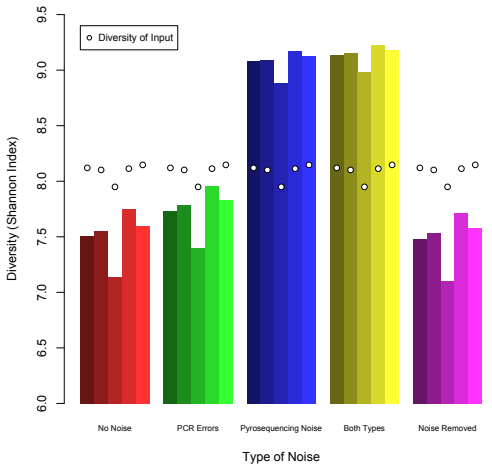

Figure 4.70: Diversity for all Group D1 output datasets (see Table 4.4). Each simulation was repeated 5 times.

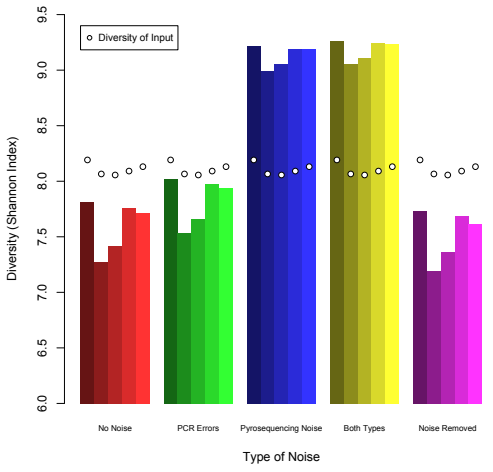

Figure 4.71: Diversity for all Group D2 output datasets (see Table 4.5). Each simulation was repeated 5 times.

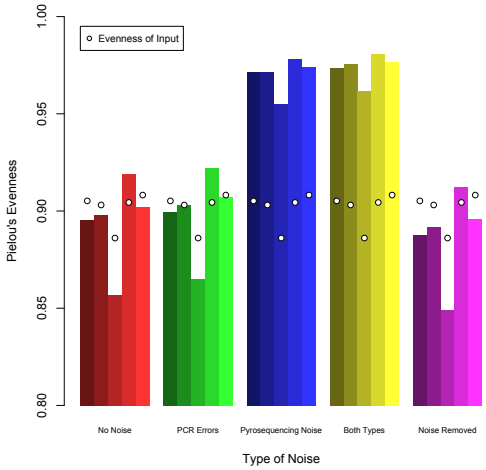

Figure 4.72: Evenness for all Group D1 output datasets (see Table 4.4). Each simulation was repeated 5 times.

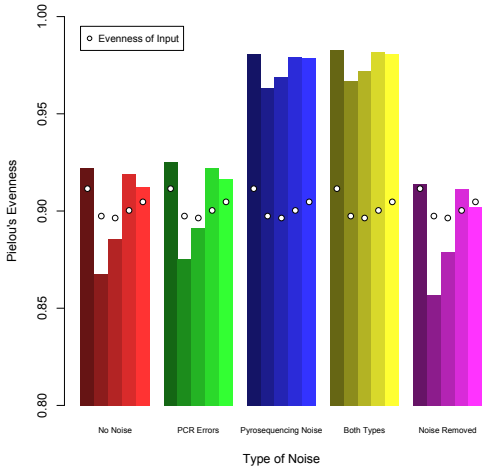

Figure 4.73: Evenness for all Group D2 output datasets (see Table 4.5). Each simulation was repeated 5 times.

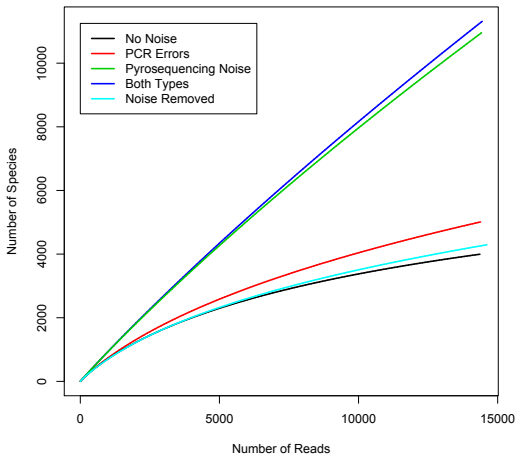

Figure 4.74: Group D1 datasets (see Table 4.4): Rarefaction curves using output data from simulations on datasets with different types of simulated noise. The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

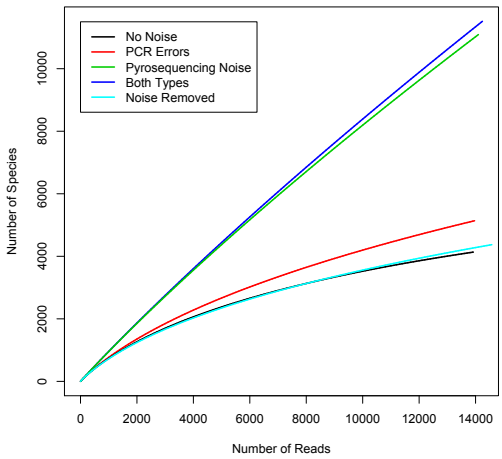

Figure 4.75: Group D2 datasets (see Table 4.5): Rarefaction curves using output data from simulations on datasets with different types of simulated noise. The curves were generated by randomly accumulating the reads, starting from one read, and counting the number of species present at each point.

4.3.7 Best Practice

Table 4.8 shows how well it can be expected that different types of dataset can be analysed. The first three columns show attributes (sample size, richness and diversity) that can be measured in the data; the following two columns show how these attributes estimate the true richness and variance of the abundance distribution; the next two columns show the effect of each type of dataset on chimera detection and chimera generation; the final column shows the recommended chimera removal method.

This table, along with Table 4.9, shows that, in general, datasets with a large number of reads but low species richness are easier to analyse and will therefore give more accurate results. Additionally, a high species diversity, indicating an abundance distribution with higher variance, is desirable for chimera detection but not for community analysis.

The UCHIME reference method is recommended for datasets with a lower sample size because the negative effects of fewer reads are less pronounced with this approach. Perseus is recommended for datasets with low richness and high diversity because it was found to outperform UCHIME on datasets of this type, as shown in Figure 4.39 in Section 4.3.4. The UCHIME *de novo* method is recommended for the remaining types of dataset because it is either better than Perseus or comparable to Perseus in terms of results but much better in terms of speed.

Before Tables 4.8 and 4.9 are consulted, it is necessary to consider the following points.

- **Noise Removal:** It is important that PCR noise and sequencing noise have been removed from the data as comprehensively as possible because the presence of noisy data severely affects the performance of chimera removal software (Section 4.3.2). For pyrosequencing data, such as those simulated in this chapter, the AmpliconNoise pipeline is an effective tool for this purpose. For Illumina data, the Illumina processing pipeline available in QIIME (21) is recommended.

- **Sample Size:** The number of reads in the dataset is a very important factor to consider. Whilst a dataset with a relatively low number of reads will generally have a smaller chimera percentage, this is overridden by the fact that chimera detection is much poorer in these datasets and estimates of their community properties are much less accurate (Section 4.3.2). Thus, a sequencing strategy that maximises the number of reads is recommended and Illumina is the most appropriate NGS platform for this purpose.

- **Reference Based Chimera Detection:** For datasets with fewer reads it has been shown that the UCHIME reference method is the most effective approach (Section 4.3.3).

However, this is only the case if an appropriate reference database is available. For 16S data it has been shown that the ChimeraSlayer and RDP classifier databases are good choices. For other types of data the use of the UCHIME reference based method should depend on the quality of the reference databases available.

Measured Properties			True Properties		Performance		
Sample Size	Richness	Diversity	Richness	Variance	Chimera Detection	Chimera Generation	Recommended Software
Low	Low	Low	Unknown	Unknown	Poor	Average	UCHIME - reference method
Low	Low	High	Unknown	High	Poor	Low	UCHIME - reference method
Low	High	Low	High	Unknown	Very poor	Low	UCHIME - reference method
Low	High	High	High	High	Very poor	Very low	UCHIME - reference method
High	Low	Low	Low	Low	Good	Very high	UCHIME - *de novo* method
High	Low	High	Low	High	Optimal	High	Perseus
High	High	Low	High	Low	Poor	High	UCHIME - *de novo* method
High	High	High	High	High	Good	Low	UCHIME - *de novo* method

Table 4.8: Table to show expected performance and recommended chimera removal strategy for datasets with different properties.

	Desired Richness	Desired Sample Size	Desired Variance of Sequence Abundances	Desired Noise
Chimera Detection	Low	Large	High	None
Chimera Generation	High	Small	High	NA
Community Analysis	Low	Large	Low	None

Table 4.9: Table to show desired attributes for data to possess in order to increase performance in three different areas.

4.4 Discussion

The analysis in the previous section has demonstrated that input data exhibiting different properties can drastically affect the composition of the output data. This can, in turn, affect the accuracy and reliability of the analysis carried out on these data. It is unfortunate that, in practice, the composition of the environmental samples to be analysed cannot be chosen in advance to any degree of accuracy. However, once the data has been examined, it can be decided what allowances must be made depending on the now evident properties of the data. The main consideration is that sample composition can't be adjusted but the findings in this chapter can be used to determine the degree of confidence with which results are treated. In general, mainly because of the relatively poor performance of chimera removal software on datasets of this type, results will contain more ambiguity than it was previously believed.

In order to effectively evaluate sequencing data it is first necessary to ascertain what the purpose of the analysis is, and therefore what properties are required of the results. Table 4.9 summarises the desirable properties for data to exhibit in order to aid three areas of analysis. For example, richer samples are beneficial if reduction of chimeras is a priority, whereas samples with low richness will produce more accurate results if the data is being used for

community analysis.

It is noticeable that low input richness and a large sample size will result in improved chimera detection but also in an increased level of chimera generation. A low level of chimera generation should usually be prioritised because prevention is better than cure. However, the only attribute that a researcher will have any significant control over is the sample size (number of reads) produced which will vary with different experimental protocols and sequencing platforms. Despite the fact that small sample sizes will generally reduce the proportion of chimeras contained within the data, a large sample size will still usually be desirable because of the greater amount information that it will provide.

For conditions to be optimal for chimera detection it is required that a large sample size be taken from a dataset with a low initial richness and an abundance distribution with high variance (high σ). All noise should also have been removed. Even when this is the case around 20% of chimeras are awarded a UCHIME score of zero, meaning that they are misclassified as good sequences regardless of the acceptance threshold. This is an extremely worrying result, especially when it is considered that conditions will usually be suboptimal (noise removal will rarely be close to perfect, for example) and, therefore, the number of undetected chimeras will generally be even greater than this figure of 20%.

One possible reason for these misclassifications is that a chimera could potentially have a higher abundance than one or both of its parents, and therefore would not be detected by UCHIME. This can happen when a chimera is generated during an early round of PCR and one of its parents subsequently experiences more instances of PCR extension failure than the chimera. Parent sequences may also have lower abundances (or be missing completely) because of the random sampling.

Closer inspection of the data shows that this is not usually the case for chimeras with UCHIME scores of zero and, therefore, there must be another reason for the presence of these misclassified chimeras. The presence of the candidate chimera, plus both of its higher abundant parents in the analysed data suggests that there could be a problem with the chimera checking software itself.

In the Introduction chapter to this thesis (Section 1.4.4) the UCHIME algorithm is presented, showing that the UCHIME score is calculated using the formulae:

$$H_L = \frac{Y_L}{\beta(N_L + n) + A_L},$$

$$H_R = \frac{Y_R}{\beta(N_R + n) + A_R}$$

and

$$H = H_L \times H_R$$

where H is the final UCHIME score, H_L and H_R are the UCHIME scores for the left and right parts of the alignment respectively, Y_L, Y_R, N_L, N_R, A_L and A_R are 'yes', 'no' and 'abstain' votes for each part of the alignment and β and n are input variables used to weight the effect of a 'no' vote.

It can be seen that, for a UCHIME score of zero to be returned, the number of 'yes' votes on one or both parts of the alignment must be equal to zero. This is not the case when a three-way alignment between a chimera and its true parents is invoked and, therefore, the problem must be due to UCHIME selecting the wrong parents from the dataset. It can be concluded that existing chimera removal software is not adequate and improvements are required in order to eradicate undetected chimeras as a significant source of noise in sequencing data.

The poor performance of chimera detection software on *in silico* datasets highlights the fact that earlier testing methods using mock communities were insufficient. There is now a big incentive to generate datasets containing realistic chimeras, simulated using algorithms such as those described in Chapter 3, with the goal of further testing and, ultimately, improving the chimera removal process.

Chapter 5

Constructing Interaction Networks using Pyrosequencing Data

5.1 Introduction

Due to the large population size and diversity found in meiofauna communities, along with the added difficulty of observing the organisms therein, relatively little is known about their ecology compared with that of communities featuring larger organisms. In this report, various ways of determining relationships between such organisms within a community are investigated and evaluated. Many of the techniques, most of which involve the use of DNA sequencing data, could also be applied for the analysis of other types of organism and it is anticipated that they will be particularly relevant for analysing microbial communities. However, all of the investigations described in this report deal with meiofauna, particularly nematode, data.

Nematodes, or roundworms, make up a particularly abundant and diverse phylum with over 28000 species described and an estimated 1 million species in total (112). Nematodes have adapted to nearly every ecosystem on Earth and are particularly abundant on the ocean floor where they are the dominant meiofaunal phylum. The data used in this report were generated from marine benthic samples from coastal regions around Europe. Because nematodes are so populous, especially in the chosen environment, there is ample opportunity to explore relationships between different species of nematode localised within one site and also across a range of sites. For these reasons, studying the relationships between nematode species, and species of other meiofauna phyla in their environment, should provide a good opportunity to yield worthwhile results regarding inter-species interactions and establish a good platform to apply the same methods to other organisms and communities.

Organisms interact with each other in a number of different ways and some of these may not have intuitively obvious effects (on abundance data, for example). For meiofauna, most interactions will be based on feeding relationships and competition for resources. The most obvious interaction is the link between predator and prey, with high abundance of a prey species having a positive impact on predator species. Indirect predator prey relationships are also possible and these are explained further in Section 5.3.2. Other possible interactions include competition, amensalism, mutualism and commensalism (113) (114).

Section 5.2.1 describes an experimental dataset containing nematode sequencing data. In this data, the presence of unexpected species in samples can be used to infer direct and indirect predator-prey relationships as described in Section 5.3.

Faust et al. (36) use co-occurrence data to infer relationships within the human microbiome. Section 5.4 describes the use of the meiofauna co-occurrence data (outlined in Section 5.2.2) to attempt the inference of interactions and several methods of doing this are discussed. The reliability of using data of this kind is debatable as a number of scenarios can be envisaged where co-occurrence of two or more species will not necessarily infer the expected relationship. This may be the case in situations where environmental factors play an important part in the determination of species distributions. In such cases, the effect of interactions on co-occurrence may be outweighed by these factors.

Comparisons between the two sets of results from the experiments in Sections 5.3 and 5.4 will produce interesting conclusions regarding the effectiveness of using co-occurrence data to infer these types of relationships.

5.2 Data

5.2.1 Dataset of Individually Sequenced Nematodes (Individual Dataset)

A series of experiments where pooled samples of multiple nematodes, were sequenced and analysed in order to investigate the formation of chimeras (46) are described in Chapter 2. In addition to this, similar experiments were carried out on 74 samples each containing a single species of nematode. The exact sequences for these 74 single nematodes were known because they had been found separately using Sanger sequencing. Each sample was identified by the primer used for sequencing, labelled P1 to P74.

The samples were sequenced, denoised and filtered for chimeras. All of the remaining sequence data in the whole dataset were clustered using a complete-linkage clustering algorithm to determine Operational Taxonomic Units (OTUs) at a level of 99% similarity. Megablast was used on the GenBank/EMBL/DDBJ nucleotide database for taxonomic assignment and the OCTUPUS annotation and parsing toolkit (49) was used for OTU annotation, this was restricted to matches of 90% or better.

Despite containing only a single nematode, each sample was revealed to contain a number of additional OTUs at lower abundance, some corresponding to those from one or more of the other nematode samples.

5.2.2 Dataset of Meiofaunal Communities (Community Dataset)

Chapter 2 presents research conducted on an experiment involving meiofauna sequencing data (45). In this study, marine benthic samples were collected from 23 sites in the UK, France, Spain, Portugal and Gambia.

Again, the samples were sequenced, denoised and filtered for chimeras and all of the remaining sequence data were combined and then clustered to find 99% OTUs which were taxonomically assigned. After taxonomic assignment, the dataset could easily be filtered for specific phyla if required; a co-occurrence dataset featuring nematode data only was produced.

5.3 Methods - Analysis of Individual Dataset

5.3.1 Food Web Construction

Using the individual nematode dataset (Section 5.2.1), the consensus sequences obtained for the 99% OTUs were combined with the 74 Sanger sequences known for each Nematode in the experiment. These sequences were aligned against each other using MAFFT (FFT-NS-2 algorithm) and FastTree was used to construct a phylogenetic tree from this alignment. Using the tree, the original Sanger sequences were matched to the OTUs that differed to them by no more than 1%, some of the Sanger sequences matched multiple OTUs and the data for these were merged together. Some of the Sanger sequences were deduced to be from the same organism, some returned no match and one of the sequences was corrupt, this reduced the number of unique successful experiments to 56.

74 Individual Nematode Samples

342 OTUs	A1	A2	A3	A4	A5	A6	A7	A8	→
B1	10	750	0	0	0	0	0	0	
B2	0	0	0	0	0	0	0	0	
B3	0	10	0	0	500	0	0	0	
B4	500	0	0	0	0	0	0	0	
B5	20	0	0	10	20	0	1000	0	
B6	0	10	0	0	0	0	0	0	
B7	0	0	0	0	0	20	0	0	
B8	10	0	0	0	0	15	0	10	
↓									

(a) To begin with the OTUs from each sample were arranged into a matrix of abundances. The abundances are the number of reads which were sequenced for each OTU. All Sanger sequences were then compared to the OTU sequences by forming a phylogenetic tree of the alignment and those differing from each other by less than 1% were identified. In this example, the Sanger sequences relating to samples A4 and A5 were found to be the same.

56 Individual Nematode Samples

268 OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
B1	10	750	0	0	0	0	0	0	
B2	0	0	0	0	0	0	0	15	
B3	0	10	0	500	0	0	0	0	
B4	500	0	0	0	0	0	0	0	
B5	20	0	0	30	0	1000	0	0	
B6	0	10	0	0	0	0	0	0	
B7	0	0	0	0	0	20	0	0	
B8	10	0	0	0	0	15	0	10	
↓									

(b) OTU and Sanger sequences were merged together by combining matching columns and rows and summing the abundances in the merged cells. The result was a smaller matrix of 268×56 - note that sample A5 has now been combined with sample A4. Sanger sequences that matched an OTU sequence are highlighted in this matrix, in all cases each OTU had a high relative abundance in the sample which verified that it did indeed represent the main nematode that was sequenced.

56 Individual Nematode Samples

56 Matching OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
A1	500	0	0	0	0	0	0	0	
A2	10	750	0	0	0	0	0	0	
A3	0	0	600	0	10	0	0	25	
A4	0	10	0	500	0	0	0	0	
A6	0	0	0	0	750	0	0	0	
A7	20	0	0	30	0	1000	0	0	
A8	0	0	0	0	0	0	500	0	
A9	0	0	0	0	0	0	0	800	
↓									

(c) The OTUs that matched the Sanger sequences were relabeled to the relevant matching sample label and the non-matching OTUs were removed. The rows were reordered so that their labels were in the same order as the column labels. The rows of the matrix now show the abundances of the nematode species which were present in each main nematode sample - these represent the prey of the main nematode.

56 Predator Nematodes

56 Prey Nematodes	A1	A2	A3	A4	A6	A7	A8	A9	→
A1	0	0	0	0	0	0	0	0	
A2	1	0	0	0	0	0	0	0	
A3	0	0	0	0	1	0	0	1	
A4	0	1	0	0	0	0	0	0	
A6	0	0	0	0	0	0	0	0	
A7	1	0	0	1	0	0	0	0	
A8	0	0	0	0	0	0	0	0	
A9	0	0	0	0	0	0	0	0	
↓									

(d) To produce a food web, the diagonal cells of the matrix were replaced with zeros because it is assumed that the nematodes have no feeding interactions within their own species. All other non-zero cells were converted to a value of one to signify a feeding interaction. In the example above it can be seen that nematode A1 preys on nematode A2 which, in turn, preys upon nematode A4.

Figure 5.1: This figure shows how data from the separate sequencing of individual nematodes can be used to generate a food web. Additional information was obtained from the Sanger sequencing of each individual nematode which provided a known sequence for each experiment. The above example is intended to be illustrative and does not portray identical data to those used for the analysis in this chapter.

The resulting matrix consisted of the 56 known nematodes and an additional 268 unknown species containing OTU abundance data where the abundances were taken to be the number of reads sequenced for each OTU. This matrix could easily be converted into a Food Web by adding a link anywhere that a non-zero read was present, ignoring interactions between a species and itself. It was, of course, not possible to identify feeding relationships between the unknown organisms, only which of the 56 known nematodes they were eaten by.

A 56×56 subset of this food web, consisting only of the feeding relationships between the known nematode species, was derived by filtering out OTUs that did not match with any one of the Sanger sequences. The methods used to generated a food web in this way are illustrated in Figure 5.1.

5.3.2 Direct and Indirect Effort Matrices

56 Individual Nematode Samples

56 Matching OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
A1	0	0	0	0	0	0	0	0	
A2	10	0	0	0	0	0	0	0	
A3	0	0	0	0	10	0	0	25	
A4	0	10	0	0	0	0	0	0	
A6	0	0	0	0	0	0	0	0	
A7	20	0	0	30	0	0	0	0	
A8	0	0	0	0	0	0	0	0	
A9	0	0	0	0	0	0	0	0	
↓	…	…	…	…	…	…	…	…	
Total	100	120	75	80	100	150	50	100	

(a) The abundances (read numbers) of prey nematodes present in each predator nematode sample were found using the methods presented in Figure 5.1 and the total abundance in each experiment was recorded.

56 Individual Nematode Samples

56 Matching OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
A1	0	0	0	0	0	0	0	0	
A2	0.1	0	0	0	0	0	0	0	
A3	0	0	0	0	0.1	0	0	0.25	
A4	0	0.083	0	0	0	0	0	0	
A6	0	0	0	0	0	0	0	0	
A7	0.2	0	0	0.375	0	0	0	0	
A8	0	0	0	0	0	0	0	0	
A9	0	0	0	0	0	0	0	0	
↓									

(b) The proportion of each prey nematode found in each nematode sample was found by dividing each column by the column total to give the effort matrix, f.

Figure 5.2: Generation of an effort matrix using OTU abundance data. The above example is intended to be illustrative and does not portray identical data to those used for the analysis in this chapter.

A matrix, f, of "efforts" has been described (115). The efforts are defined as the proportion of each predator's diet that is made up of each prey species. The f matrix was estimated from the food web by using the OTU abundance for each prey species j of predator species i to calculate the proportion f_{ij} as shown in Figure 5.2. The abundance is taken to be the number of reads sequenced for each OTU.

The corresponding matrix of all direct and indirect predator-prey relationships, F, was calculated using the formula $F = (1 - f)^{-1} f$. This was used to find the matrix of indirect interactions only, I, where:

$$I_{ij} = \begin{cases} F_{ij} & \text{if } f_{ij} = 0, \\ 0 & \text{if } f_{ij} > 0. \end{cases}$$

5.3.3 Competition Matrix

Species that consume the same resources will be in competition with each other. A measure of the level of competition between two predator species, i and j, for a particular prey species, k, can be calculated by taking the product of their (direct and indirect) predator-prey interactions for that species, $F_{ik}F_{jk}$ using data from the F matrix described in Section 5.3.2. Thus, the greater the contribution from k to the diet of either species, the greater the competition between them. The total competition between species i and species j can then be found by summing over all species k to give the formula,

$$c_{ij} = \sum_k F_{ik}F_{jk}.$$

5.3.4 OTU classification and Inferring the Diet of Nematodes

Every OTU in the individual nematode dataset was taxonomically classified. A Megablast alignment was carried out with the fasta file containing the OTU sequences aligned against the Silvamod reference database (116). The output was analysed using LCA Classifier (117) to taxonomically classify the OTUs as accurately as possible based on their sequences.

Using this classification data, the lists of all OTUs present in each individual experiment could be grouped at a specified taxonomic level (e.g. phyla). These lists correspond to the diet of the main nematode species in each individual experiment. New datasets showing this information were created by normalising or rarefying the data. The steps used for inferring the diets of the nematodes are illustrated in Figure 5.3

5.3.5 Assigning Feeding Types to Nematodes

Video data (archived at Bangor University) exists for each single nematode that was sequenced but, unfortunately, there is no footage of the nematodes actually eating each other. However, through expert analysis of the worms' morphology, conducted by a group led by Tom Moens at Ghent University, it has been possible to determine the feeding types of a selection of the nematodes. A total of 25 nematodes were classified by their feeding types according to Wieser (118) and according to Moens and Vincx (119). Information was also included regarding their presumed main food sources and potential secondary food sources.The Wieser feeding types are expanded upon below.

- **1A - Selective deposit feeder.** Minute mouth, hence probably bacteria as main food.

56 Individual Nematode Samples

268 OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
B1	10	750	0	0	0	0	0	0	
B2	0	0	0	0	0	0	0	15	
B3	0	10	0	500	0	0	0	0	
B4	500	0	0	0	0	0	0	0	
B5	20	0	0	30	0	1000	0	0	
B6	0	10	0	0	0	0	0	0	
B7	0	0	0	0	20	0	10	0	
B8	10	0	0	0	15	0	0	0	
↓									

(a) OTUs from each sample were arranged into a matrix of abundances. The abundances are the number of reads which were sequenced for each OTU. All Sanger sequences were then compared to the OTU sequences by forming a phylogenetic tree of the alignment and those differing from each other by less than 1% were identified. Sanger sequences that matched an OTU sequence are highlighted in this matrix.

56 Individual Nematode Samples

268 OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
B1	10	0	0	0	0	0	0	0	
B2	0	0	0	0	0	0	0	15	
B3	0	10	0	0	0	0	0	0	
B4	0	0	0	0	0	0	0	0	
B5	20	0	0	30	0	0	0	0	
B6	0	10	0	0	0	0	0	0	
B7	0	0	0	0	20	0	10	0	
B8	10	0	0	0	15	0	0	0	
↓									

(b) Cells of the matrix marking the intersection of each sample and its matching OTU were replaced with zeros because it is assumed that the nematodes have no feeding interactions within their own species.

56 Individual Nematode Samples

268 OTUs	A1	A2	A3	A4	A6	A7	A8	A9	→
P1	10	0	0	0	0	0	0	0	
P2	0	0	0	0	0	0	0	15	
P1	0	10	0	0	0	0	0	0	
P1	0	0	0	0	0	0	0	0	
P1	20	0	0	30	0	0	0	0	
P3	0	10	0	0	0	0	0	0	
P4	0	0	0	0	20	0	10	0	
P3	10	0	0	0	15	0	0	0	
↓									

(c) Each OTU was classified using LCA classifier and the OTUs were labelled with their respective phyla. Here, different colours represent different phyla.

56 Individual Nematode Samples

33 Unique Phyla (Diet)	A1	A2	A3	A4	A6	A7	A8	A9	→
P1	110	75	0	90	10	0	20	0	
P2	0	0	0	0	0	0	10	45	
P3	20	10	0	0	25	0	0	0	
P4	0	0	0	0	50	0	10	0	
P5	0	0	0	0	0	100	40	0	
P6	0	20	30	0	0	0	0	0	
P7	0	0	20	0	0	0	15	0	
P8	0	0	35	0	10	0	0	0	
↓									

(d) Rows corresponding to each of the 33 different phyla were combined and the abundances were summed together. This gives a representation of the diet of each nematode. The abundances can be divided by the total abundance in each column to give the proportion of each nematode's diet that is composed of each phylum.

Figure 5.3: This figure shows how data from the separate sequencing of individual nematodes can be used to infer the diet, in terms of the different phyla ingested, for each nematode. Additional information was obtained from the Sanger sequencing of each individual nematode which provided a known sequence for each experiment. The above example is intended to be illustrative and does not portray identical data to those used for the analysis in this chapter.

- **1B - Non-selective deposit feeder.** May feed predominantly on microalgae or on bacteria; role of protists as food poorly known.
- **2A - Epistratum feeder.** Many of these appear to feed on diatoms and other microalgae, but this may not be their only food.
- **2B - Predator/Omnivore.**

To test its accuracy, these feeding types could be compared with the trophic levels from the food web. It is also possible that there could be a relationship between the diversity of a nematode's diet and its feeding type - a deposit feeder, for example, may appear to have eaten a wide variety of "prey" because of the range of dead organic matter that it has consumed. The Shannon diversity index was therefore used to calculate the diversity of phyla in each nematode's diet and also the nucleotide diversity of the OTU sequences in each diet. More evidence that the data from the individual nematode experiment are food web data would be obtained if the differences between inferred diets of nematode species with the same feeding type were smaller than those of species with different feeding types.

5.3.6 NMDS Analysis

Nonmetric multidimensional scaling (NMDS) is a technique used in community analysis to help visualise differences between various sites in terms of their taxonomic composition. As the taxa richness increases in the sites it becomes increasingly difficult to compare them graphically due to the high dimensionality, NMDS attempts to rectify this issue by reducing the data to two or three dimensions which still retain the same information.

NMDS is initiated using a distance or dissimilarity matrix such as a matrix of Bray-Curtis dissimilarities. An iterative procedure is then applied which uses the ranking of taxa abundances at each site.

- An initial two dimensional plot is formed and regressions of the distances on this and the measured distances are carried out.
- The difference (or "stress") between the predicted values from the regression and the 2D configuration is determined. If the 2D configuration kept the ranked abundances in the original order then a plot of one against the other would be monotonically increasing. The level of stress is determined by how much the 2D points differ from this relationship.
- The 2D points are positioned in a configuration which reduces the stress.

- The procedure is repeated for a set number of iterations or until the stress is below a certain threshold. Any value below 0.2 is usually taken as indication that the two dimensional plot gives a very good representation of the data.

NDMS was applied to the individual members that had been assigned by feeding type using the *metaMDS* function in the *Vegan* package in **R**. The abundance data for each phyla present in the samples were used for the community data with each individual nematode analogous to a "site". Thus, the differences in the nematode's diets could be analysed using this method.

5.3.7 Permutation ANOVA

A test was required to see if there were significant differences in the composition of nematodes of each feeding type. The *adonis* function in the *Vegan* package in **R** was used to perform a permutation analysis of variance on the inferred diets of the nematodes that had been categorised by feeding type. This was repeated when only two feeding types were defined (combining 1A with 1B and 2A with 2B). It was also repeated when the nematode diets were rarefied to the same abundance of the sample with the lowest OTU abundance.

The *adonis* function performs a partition multivariate analysis of variance which partitions distance matrices among sources of variation and performs permutation tests to determine the significance of the partitions. In this case Bray-Curtis distances were calculated against only one partition, the feeding type. The permutation tests work by generating 999 random permutations of the observed data and performing ANOVA on each of these. The F-statistics returned from these tests are compared with the F-statistics returned from an ANOVA test on the true data to calculate the p-values which determine the significance levels.

5.3.8 Multinomial Logistic Regression

Multinomial logistic regression is an extension of standard logistic regression in which the response variable is categorical but can take more than two values. This method can be used to determine the feeding type (1A, 1B, 2A or 2B) of a newly observed organism based solely on its inferred diet. Multinomial logistic regression can be carried out in **R** using the *multinom* and *predict* functions to return the probabilities of new observations belonging to each feeding type.

5.4 Methods - Analysis of Community Dataset

5.4.1 Pre-processing and Post-processing the Data

Before interactions could be inferred from the co-occurrence data in the community dataset (Section 5.2.2) it was necessary for the following pre-processing to take place:

- The abundance data were normalised so that the abundances at each site summed to the same total.
- All OTUs with fewer than three reads were filtered out.
- Only OTUs with an abundance greater than the median abundance were included.
- Very dominant OTUs (for example those with greater than 60% relative abundance) that showed little change across samples were to be removed, however no such OTUs were found to exist.
- A Hellinger transformation was applied to the data to counteract any additional dominance effect in the sample. No transformation of the data was used in conjunction with the SparCC analysis because the SparCC software uses a Bayesian approach to transform the data (40).

After the interaction networks were inferred, Benjamini-Hochberg corrections were applied to the data to account for false positives using the *p.adjust* function in **R** and significant p-values ($p < 0.05$) were used to determine interactions between OTUs as per the methods used by Berry and Widder (120).

5.4.2 SparCC

The SparCC (40) approach for inferring interactions between species involves the estimation of linear Pearson correlations between the log-transformed ratios,

$$y_{ij} = \ln \frac{x_i}{x_j}$$

where x_i is the proportion of OTU i and x_j is the proportion of OTU j. The use of these ratios is the defining feature of SparCC because y_{ij} is equal to the ratio of the true abundances of OTUs i and j and, critically, because y_{ij} is independent of the abundances of all other OTUs in the dataset. It is assumed that the number of OTUs is large and that the resultant matrix will be *sparse*, meaning that the majority of OTU pairings will have no significant interaction.

Variances of y_{ij},

$$t_{ij} = \text{Var}(y_{ij}),$$

can be calculated across all sites to give an indication of relationships between OTUs. Perfectly correlated OTU pairs will yield constant values for y_{ij} and, therefore, have zero variance. Conversely, uncorrelated OTU pairs will produce high values for t_{ij}. Following on from this, a set of equations can be formulated to relate t_{ij} to the correlation, ρ_{ij}, between the true abundance of the OTUs:

$$t_{ij} = \omega_i^2 + \omega_j^2 - 2\rho_{ij}\omega_i\omega_j$$

where ω_i^2 and ω_j^2 are the variances of the log-transformed true abundances of OTUs i and j respectively. It is apparent that if t_{ij} is less than the sum of the two variances then the correlation is positive and if t_{ij} is greater than the sum of the two variances then the correlation is negative. These equation are not solvable analytically but the values for ρ_{ij} can be estimated using an iterative procedure.

First, estimates for the variances of the true abundances, ω_i for all i, are calculated using the assumption that all OTUs are uncorrelated. These values can then be used in conjunction with the values for t_{ij} to give estimates for ρ_{ij}. Accuracy is improved by repeating this step, with the OTU pair showing the strongest correlation at each step omitted from the true variance calculations in subsequent steps, until all OTUs have been removed.

5.4.3 L1 Penalised Sparse Regression Model

This section discusses a method of inferring interactions from co-occurrence data involving the calculation of a precision matrix, which is simply the inverse of the covariance matrix. If it can be assumed that the majority of species have negligible relationships between each other then a sparse precision matrix is a sensible option to use to attempt to model the interactions. This is a matrix where most of the parameters are estimated as zero and the remaining non-zero parameters indicate positive or negative interactions, depending on their sign.

Banerjee et al (121) describe algorithms to estimate the precision parameters of a Gaussian distribution given the constraint that the precision matrix is sparse, the approach used is to solve a maximum likelihood problem with an added l_1-norm penalty term. In order to apply the algorithm to the meiofauna community dataset (nematode data only) the data were transformed so that a Gaussian distribution would be an appropriate model - the natural

logarithms of the relative abundance data were calculated,

$$y_{ij} = \log\left(\frac{x_{ij}+1}{\sum_i x_{ij}+s}\right),$$

where x_{ij} is the abundance of OTU i at site j and s is the total number of OTUs. The precision matrix was estimated from these transformed data using the block coordinate descent algorithm presented by (121).

5.4.4 Correlation and Dissimilarity Matrices

Four methods of calculating relationships directly from the nematode community data are the Pearson (37) and Spearman (38) correlation coefficients, the Kullback-Leibler divergence (39) and the Bray-Curtis dissimilarity (34). These were the four measures used to produce the ensemble score based network in (36) and their formulae are shown below in the context of analysing co-occurrence data.

It should be noted that often in ecology these measures are used to find similarities and differences between sites/samples in terms of their OTU composition. For the analysis in this chapter, however, they are used to find similarities and differences between OTUs depending on their distribution across different samples. The following definitions reflect their use in this context.

Pearson Correlation Coefficient

The Pearson correlation coefficient is

$$r_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)\left(\frac{Y_i-\bar{Y}}{s_Y}\right)$$

where n is the overall number of samples in the dataset, X_i and Y_i are the respective abundances of OTUs X and Y in sample i, $\bar{X}$ and $\bar{Y}$ are the respective mean abundances of OTUs X and Y in sample i and s_X and s_Y are the standard deviations of the abundances of X and Y across all samples.

Pearson's coefficient is the most commonly used measure of correlation between sets of data. It can also be rewritten as the covariance of the two OTUs divided by the product of the OTU standard deviations. This gives a value of 0 if the two OTUs are uncorrelated, a value of +1 if there is total positive correlation and a value of -1 if there is total negative

correlation.

Spearman Correlation Coefficient

The Spearman correlation coefficient is the Pearson correlation coefficient between the ranked variables. The samples for each OTU are ranked in order of abundance, the least abundant sample is assigned the rank 1, and so on. Samples with the same abundance are assigned the mean of what their collective ranks would be if ties were broken randomly. The coefficient is

$$\rho_{XY} = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}$$

where n is the overall number of samples in the dataset and x_i and y_i are the ranks of sample i for OTUs X and Y respectively.

Clearly, Spearman's coefficient is useful when differences in values between variables in a dataset are deemed to be of lesser importance than their rank order. For the analysis of co-occurrence data, different information can be gained from the two correlation coefficients. The Spearman correlation will show whether the distribution of two OTUs is similar, whereas the the Pearson correlation will be more sensitive to OTU abundance counts.

Kullback-Leibler Divergence

The Kullback-Leibler divergence of Q from P is

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(i)\ln\frac{P(i)}{Q(i)}$$

where P and Q are discrete probability distributions. In the case of OTU co-occurrence data, $P(i)$ is the proportion of the abundance of OTU P that is found in sample i and similarly for Q.

The Kullback-Leibler divergence attempts to quantify the inefficiency of using one distribution in place of another. Note that this measure is a divergence rather than a distance or a metric. This means that the Kullback-Leibler divergence is asymmetric, i.e. the divergence of Q from P is different to the divergence of P from Q.

Bray-Curtis Dissimilarity Index

As has been stated in earlier chapters, the Bray-Curtis dissimilarity index is

$$B_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

where, for use in this chapter, C_{ij} is taken to be the sum of the lowest of the two abundances for samples in which both OTUs i and j occur. S_i and S_j are the total abundances of OTU i and OTU j.

Unlike the Kullback-Leibler divergence, the Bray-Curtis dissimilarity index *is* symmetric. However, it is not a distance or a metric because it does not satisfy the triangle inequality. That is, $B_{ij} + B_{ik}$ is not always greater than B_{jk}.

The Bray-Curtis index is used extensively in ecology for analysing community data. While the Pearson and Spearman correlation coefficients may be heavily influenced by big differences in the abundances of a small number of OTUs between sites, the Bray-Curtis index is more robust and allows for some variance which may be expected to occur in data of this type.

5.4.5 Evolutionary Distance Matrix

It could be supposed that interactions between species depend on how closely or distantly related they are to each other. Closely related species are perhaps more likely to share habitats suited to their similar characteristics, suggesting a positive relationship between the two species. That is, if one is present in a community then the other is more likely to also be present. However, similar species are also more likely to be competing for the same resources, suggesting a negative relationship.

These potential interactions can be explored by examining a phylogenetic tree and the simplest approach to begin this analysis would be to look at the evolutionary distances on the tree. This method was implemented using the nematode OTUs from the community dataset described in Section 5.2.2. A phylogenetic tree was produced, again using FastTree (122), from a multiway alignment of all of the OTUs. This tree, in Newick format, was analysed using the *evol.distinct* function in the *picante* package in **R** and a matrix of pairwise phylogenetic distances was found.

5.5 Methods - Analysis of Results from Both Datasets

5.5.1 Matching OTUs Occurring in Both Datasets

To compare the different methods of inferring interaction networks, it was necessary to identify the OTUs that occurred in both datasets. To do this, the same method as described in Section 5.3.1 was used. The OTUs from both datasets were combined and aligned, and a tree was produced from this alignment using FastTree. OTUs from opposing datasets were matched with one another if they differed by less than 1%. From this information, the OTUs occurring in both the 56 species Food Web described in Section 5.3.1 and the various matrices used to analyse the community data described in Section 5.2.2 could be identified, 36 OTUs in total. The food web, f matrix, and other matrices found using either dataset were filtered to remove the unwanted OTUs, transforming each into a 36×36 matrix.

5.5.2 ROC Analysis

The analysis of ROC curves here differs slightly to the methods used in Section 4.2.11 because of differences in the data. In the previous chapter, the true values were known so the false positives in the data could be identified with perfect accuracy. For the data in this section there are several interaction matrices that are attempting to predict the same things and it is necessary to choose one of these matrices to be the "gold standard" which is assumed to have predicted every interaction perfectly. Thus, if analysis yields high AUROC (area under ROC curve) values for a method which predicts interactions using a different dataset to that used by the gold standard method then this will provide evidence that both methods are successful at predicting interactions.

The f predator-prey matrix was first used to produce the gold standard with a non-zero value of f_{ij} or f_{ji} corresponding to an interaction between species i and species j and a zero value corresponding to no interaction. This was repeated using the matrix of indirect efforts, I, to produce the gold standard and again, using the L1 precision matrix and SparCC matrix. To test the data against these gold standards, the threshold that indicated an interaction was gradually incremented and was compared with the strength of the correlation, dissimilarity or divergence value in the matrix being analysed. If the value exceeded the threshold then an interaction was predicted and these predicted interactions were compared against the gold standard to determine the percentage of false positives at each threshold level.

5.5.3 Jaccard Distance

The *Jaccard distance* (123) is used in this chapter to compare interaction networks generated by different methods. A simple formula is used to show the ratio of matching edges in two graphs to the number of mismatching edges. The Jaccard distance, d_J, is given by

$$d_J = \frac{M_{01} + M_{10}}{M_{01} + M_{10} + M_{11}}$$

where M_{01} is the number of instances for which an edge is absent from the first graph but present in the second graph, M_{10} is the number of instances for which an edge is present in the first graph but absent from the second graph and M_{11} is the number of instances for which an edge is present in both graphs. Note that the quantity M_{00}, the number of instances for which an edge is absent from both graphs, is not required for the calculation of d_J.

5.5.4 Structure of Graphs

Another way to compare two different interaction network graphs is to observe the properties at corresponding nodes in each graph. Such properties investigated in this chapter are the degree of a node, a graph's clustering coefficient, the betweenness centrality of a node and the closeness centrality of a node. These properties were all calculated using the *igraph* (124) package in **R**.

Degree of a Node

The *degree* of a node is the number of edges connecting it to other nodes in the graph.

The Clustering Coefficient

The *clustering coefficient* of a graph, sometimes called the *transitivity*, demonstrates the level to which the nodes of a graph group together. It is calculated by dividing the number of closed triplets (groups of three nodes which are all connected by three edges) in a graph by the total number of connected triplets (groups of three nodes which are connected by either two or three edges).

Betweenness Centrality

Betweenness centrality is a measure of how frequently a node appears in the shortest path between two other nodes in a graph. The betweenness centrality of a node, therefore, highlights

the importance of a node in terms of its contribution to indirect interactions in an interaction network. Betweenness centrality for node v, $C_B(v)$ is calculated using the formula,

$$C_B(v) = \sum_{x,y} \frac{\sigma_{xy}(v)}{\sigma_{xy}}; x, y \neq v$$

where σ_{xy} is the number of shortest paths between nodes x and y and $\sigma_{xy}(v)$ is the number of these paths which visit node v.

Closeness Centrality

The *closeness centrality* of a node is a measure of how close the node is to all other nodes in the graph. The closeness centrality of node v, $C_C(v)$, is the reciprocal of the sum of all of the shortest distances between the node in question and all other nodes in the graph, as shown:

$$C_C(v) = \sum_{x} \frac{1}{d(x, v)}; x \neq v$$

where $d(x, v)$ is the length of the shortest path between nodes x and v.

5.6 Results - Taxonomic Classification Statistics

All of the 342 OTUs in the individual dataset were classified and the statistics are shown in Tables 5.1 and 5.2. It was possible to classify over 90% of the OTUs by phylum, almost 50% were classified by family and 12% were completely classified, with their species identifiable. Around one third of the OTUs in the dataset were nematodes and a further 17.5% were fungi. The remaining OTUs were shared between various meiofauna, protist and some plant and vertebrate groups, possibly originating from dead seaweed and fish.

Level of Classification	Count (342 Total)	Share (%)
Domain	332	97.08
Phylum	318	92.98
Class	300	87.72
Order	257	75.15
Family	165	48.25
Genus	99	28.95
Species	42	12.28

Table 5.1: OTUs in the individual dataset which were successfully classified at each level. LCA Classifier was used for classification analysis.

Statistics relating to the classification of only the main nematode, identified by its Sanger

Kingdom:Phylum	Count (342 Total)	Share (%)
Metazoa:Nematoda	113	33.04
Metazoa:Platyhelminthes	15	4.39
Metazoa:Arthropoda	12	3.51
Metazoa:Other Phyla	25	7.31
Fungi	60	17.54
Stramenopiles	27	7.89
Alveolata	18	5.26
Viridiplantae	30	8.77
Other Kingdoms	18	5.26
Unclassified	24	7.02

Table 5.2: Kingdoms and phyla of classified OTUs in the individual dataset. LCA Classifier was used for classification analysis.

sequence, in each individual experiment are shown in Table 5.3. 73 of the 74 individuals were classified because one was removed due to a corrupt sequence. Of the 73 remaining, all but one were successfully classified as nematodes, the other returning an unclassified result. The nematodes were generally well classified (more than 50%) up to family level and 8 of the 73 (11%) were completely classified. The full classifications of all 73 nematodes, cross-referenced with their unique IDs, can be seen in Appendix A.

Level of Classification	Count (73 Total)	Share (%)
Domain	73	100.00
Phylum	72	98.63
Class	72	98.63
Order	69	94.52
Family	41	56.16
Genus	24	32.88
Species	8	10.96

Table 5.3: Sanger sequences corresponding to the main nematodes in the individual dataset which were successfully classified at each level. LCA Classifier was used for classification analysis.

5.7 Results - Visualisation of Networks

Figure 5.4 shows the food web predicted from the individual nematode data after the number of nematodes was filtered down to 56. Out of the 56 nematodes, 12 predators (non-prey), 25 intermediate feeders (both predator and prey), 12 prey species (non-predator) and 7 species that had no interactions were predicted. The food web is not a *full clique* - that is, there is not an edge between all pairs of nodes on the graph.

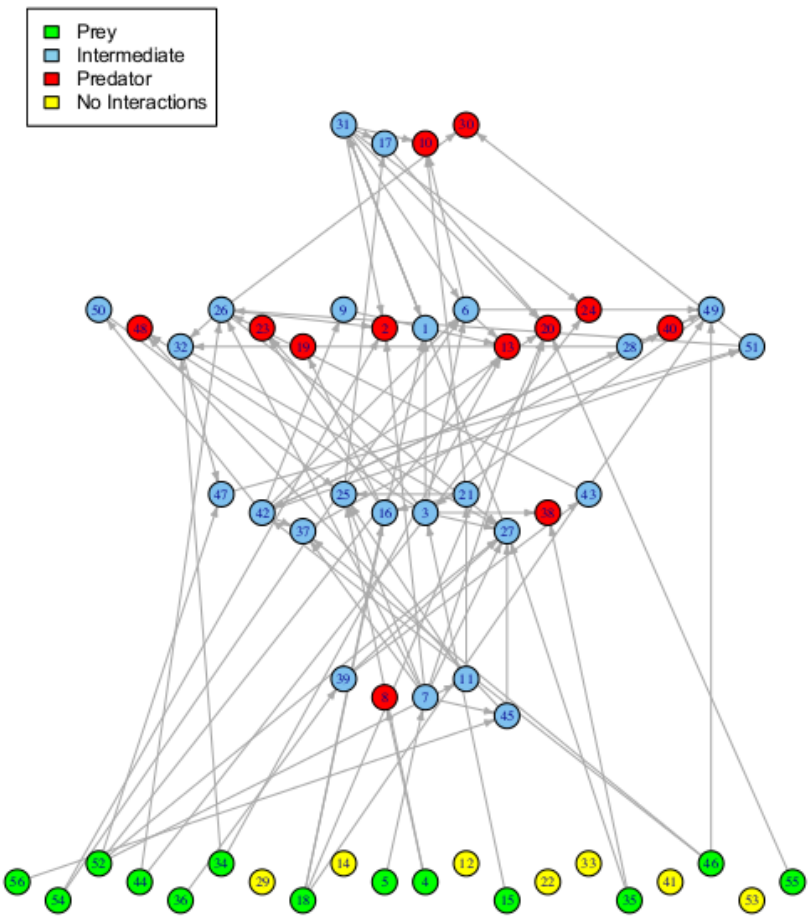

Figure 5.4: Food web of 56 nematode species. Of the OTUs generated from the 74 experiments on individual nematodes, 56 of these OTUs were found to correspond to the original Sanger sequenced nematodes. A feeding relationship between two nematodes was inferred when one of these OTUs (prey) was present in the experiment corresponding to the known Sanger sequenced nematode (predator).

5.8 Results - Comparing Feeding Types With Experimental Data

The feeding types for the 25 nematodes that were successfully categorised from the video data are shown in Table 5.4 along with their trophic levels, calculated from the food web shown in Figure 5.4. Table 5.4 also includes the Shannon diversity of the predicted prey of each nematode (prey has been categorised both by phyla and by OTU) and the nucleotide diversity of the predicted prey, also found using the Shannon index.

ID	Wieser	Moens and Vincx	Trophic Level	Diversity (Phyla)	Diversity (OTUs)	Nucleotide Diversity
P2	1B	deposit feeder	4.002	1.269	1.733	0.180
P4	2A	epistrate feeder	3.297	0.098	0.360	0.197
P7	2A	epistrate feeder	1.000	0.639	0.689	0.177
P8	2A	epistrate feeder	1.000	1.330	1.550	0.155
P9	2A	epistrate feeder	4.076	0.846	1.397	0.095
P12	2A	epistrate feeder	4.297	1.149	1.303	0.058
P13	1B	deposit feeder	4.929	1.456	1.992	0.036
P17	1B	deposit feeder	1.000	1.040	1.332	0.129
P22	2B	predator/scavenger	1.000	0.000	0.000	NA
P23	2B	predator	4.000	1.747	2.340	0.148
P26	1B	deposit feeder	1.000	1.362	2.658	0.155
P30	2A	epistrate feeder	3.597	1.182	2.211	0.133
P31	2B	predator/scavenger	3.420	1.103	2.316	0.126
P37	2A	epistrate feeder	1.000	0.315	0.703	0.041
P39	2A	epistrate feeder	1.000	0.562	1.040	0.172
P40	2A	epistrate feeder	1.000	0.637	1.099	0.000
P43	1A	microvore	3.148	0.435	0.834	0.117
P44	2A	epistrate feeder	2.000	0.566	0.645	0.093
P45	1B	deposit feeder	4.099	0.918	1.098	0.143
P47	2B	predator/scavenger	3.297	1.321	1.560	0.155
P48	2A	epistrate feeder	3.000	0.693	1.099	0.000
P51	2B	facultative predator	1.000	0.000	0.693	0.000
P62	2A	epistrate feeder	3.593	1.457	2.221	0.124
P71	2A	epistrate feeder	1.000	0.000	0.000	NA
P72	1B	deposit feeder	3.000	1.213	1.733	0.138

Table 5.4: Feeding types of nematodes compared with results from experimental data. Diversity is measured using the Shannon index.

Of the 25 categorised Nematodes, 13 of them have Wieser feeding type 2A, 6 have type 1B, 5 have type 2B and 1 has type 1A. ANOVA analyses using trophic level, phyla diversity, OTU diversity, nucleotide diversity and all combinations thereof as explanatory variables to predict feeding type were carried out, however none of these returned a significant p-value.

Figure 5.5 shows the inferred diet of the nematode species, *Chromadorita tentabundum* (ID P9) which appears to predominantly consume platyhelminthes and other nematodes, with a small fraction of the diet being comprised of fungi. Analysis of the morphology of the nematode suggested that *Chromadorita tentabundum* has feeding type 2A (epistratum feeder).

Similar diagrams can be formed for all nematodes in the individual dataset but only those for the 25 nematodes that were categorised by feeding type can be verified for accuracy.

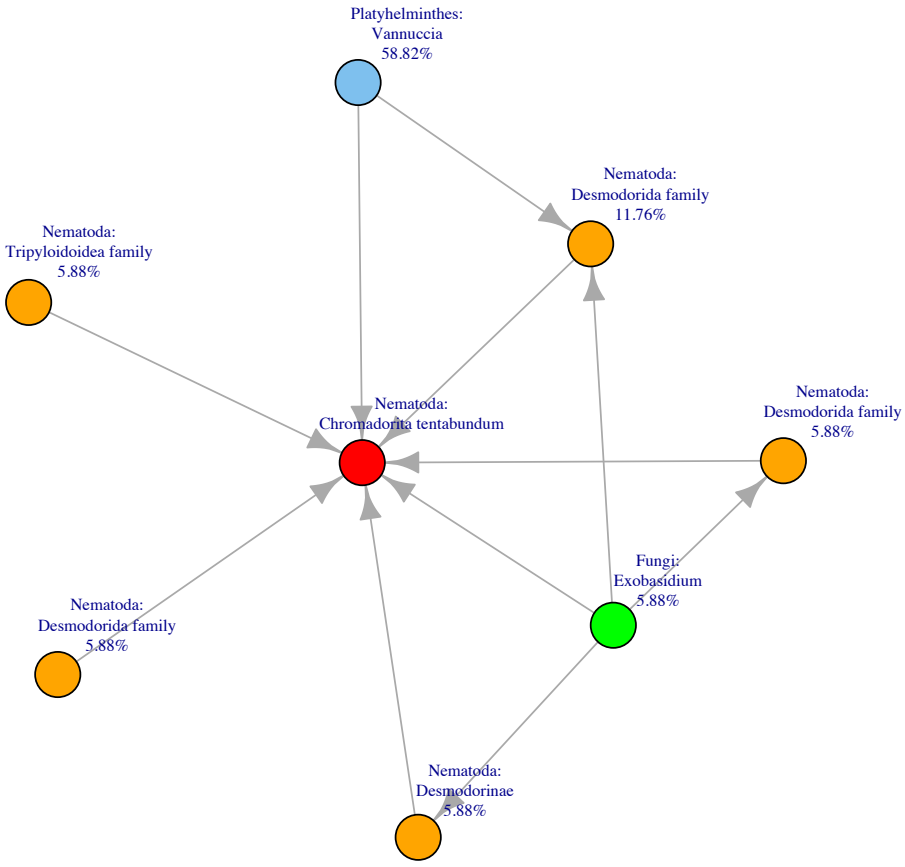

Figure 5.5: Organisms consumed by *Nematoda:Chromadorita tentabundum*. The predator was identified by classifying the OTU corresponding to the original Sanger sequenced nematode in one of the 74 experiments and prey was found by classifying all other OTUs generated from this experiment. LCA Classifier was used for classification analysis.

The bar charts in Figures 5.6 to 5.9 show the composition of the inferred diets for each Wieser feeding type based on normalised data with the OTUs grouped by phyla. The four feeding types differ in composition with the type 1A nematodes predominantly consuming Streptophyta, type 1B nematodes mainly consuming Dikarya and Platyhelminthes, type 2A nematodes consuming a wider variety of organisms with Nematoda the major dietary component and type 2B nematodes primarily consuming other nematodes.

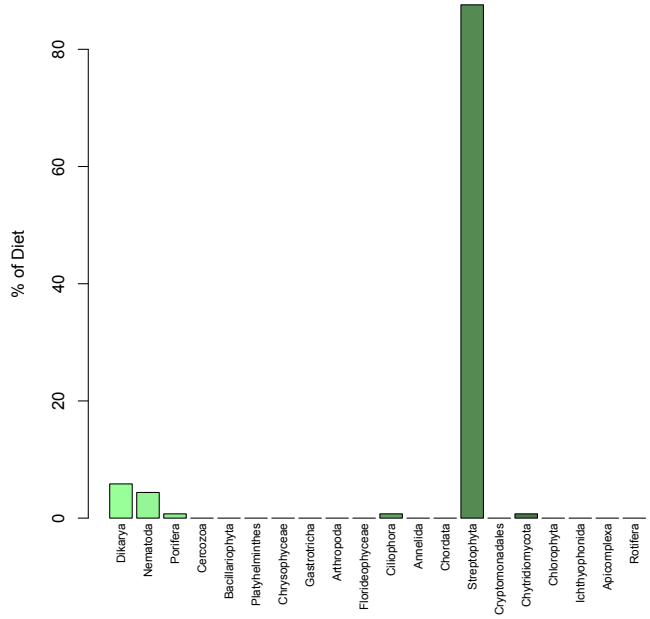

Figure 5.6: Inferred diet of Wieser feeding type 1A nematodes based on normalised data. Nematodes were categorised from videos based on their morphology and their diet was found by classifying all OTUs generated from the corresponding experiment. LCA Classifier was used for classification analysis.

The NMDS plot in Figure 5.10 returned a stress factor of 0.137 suggesting that it gives a very good representation of the multidimensional data. The plot shows how closely related the individual nematodes are to each other based on the Bray-Curtis dissimilarities calculated from their inferred diets after the data had been normalised. There appears to be a degree of grouping between the type 1B nematodes and between the type 2B nematodes. The type 2A nematodes are more spread out while the solitary type 1A nematode is isolated as distinct from the rest. Figure 5.11 shows the same plot but with ellipses drawn to depict the spread of each feeding type. It can be seen that nematodes with feeding types 1B and 2B seem to have more specific diets and have a small intersection with each other whereas nematodes of type 2A have a much more varied diet from individual to individual.

The fungal phyla Dikarya is known to contain *nematophagous fungi* of the order Basidiomycetes (125) which are parasitic fungi that are known to specifically attack nematodes. Because of the limitations of classifying the OTUs in the individual nematode dataset (due to their sequence lengths and composition), it was not possible to verify which of the Dikarya OTUs were nematophagous. However, many were classified as such up to the order level, thus it was reasonable to believe that they were parasites and, therefore, not part of the nematodes' diets.

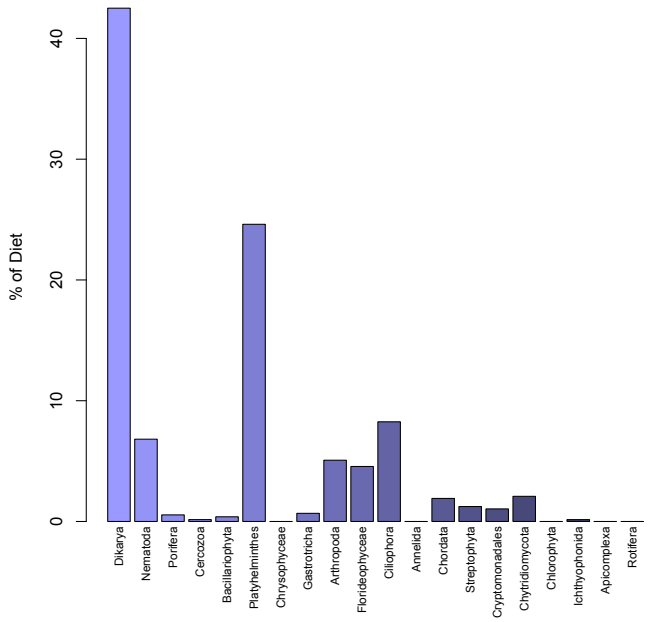

Figure 5.7: Inferred diet of Wieser feeding type 1B nematodes based on normalised data. Nematodes were categorised from videos based on their morphology and their diet was found by classifying all OTUs generated from the corresponding experiment. LCA Classifier was used for classification analysis.

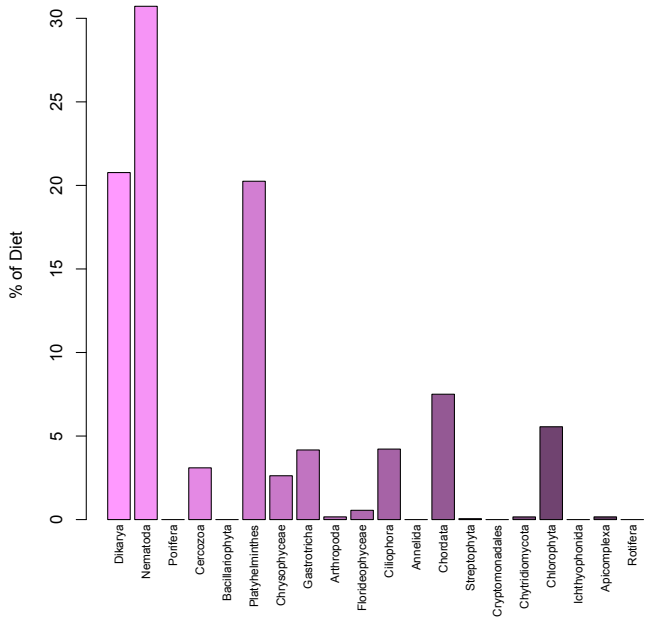

Figure 5.8: Inferred diet of Wieser feeding type 2A nematodes based on normalised data. Nematodes were categorised from videos based on their morphology and their diet was found by classifying all OTUs generated from the corresponding experiment. LCA Classifier was used for classification analysis.

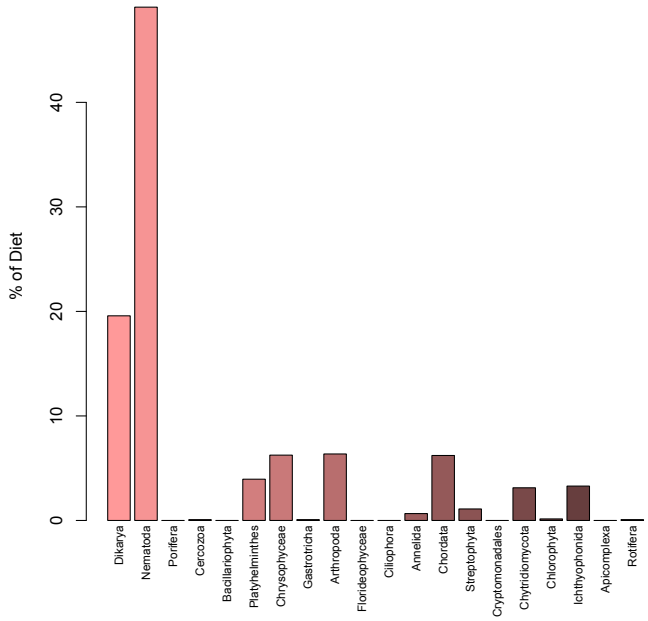

Figure 5.9: Inferred diet of Wieser feeding type 2B nematodes based on normalised data. Nematodes were categorised from videos based on their morphology and their diet was found by classifying all OTUs generated from the corresponding experiment. LCA Classifier was used for classification analysis.

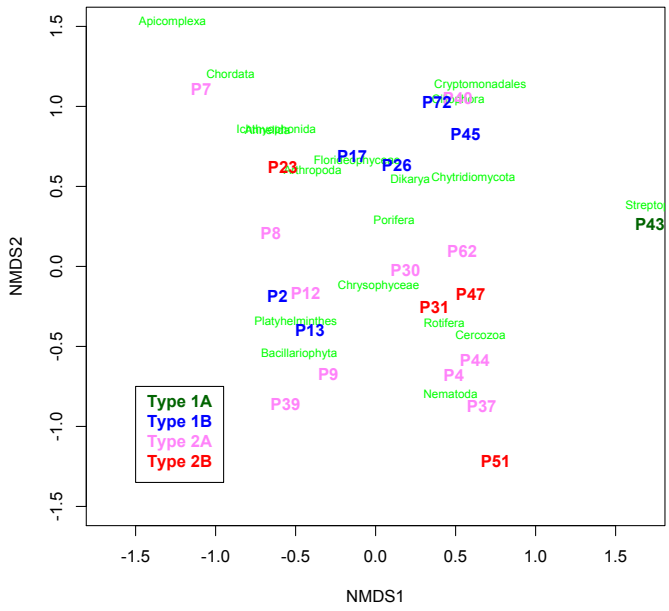

Figure 5.10: NMDS plot for individual nematodes based on Bray-Curtis distance calculated using their normalised inferred diets. Points representing nematodes are plotted close to points representing their diet.

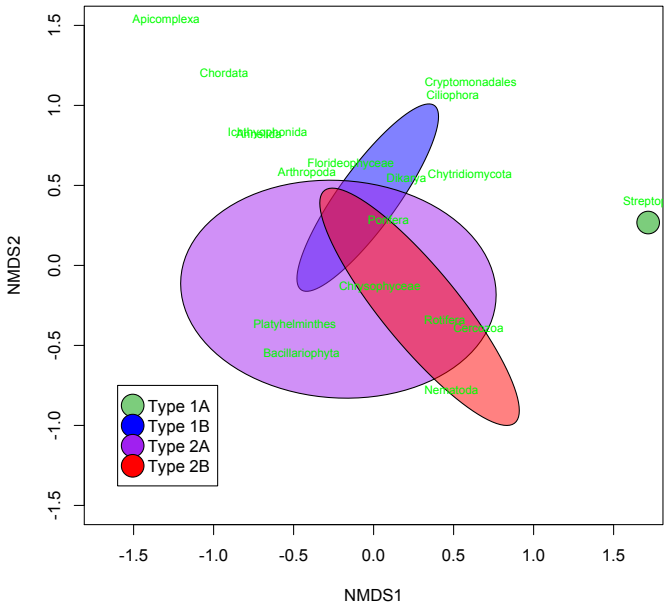

Figure 5.11: NMDS plot for individual nematodes showing the spread of nematodes of each Wieser feeding type. The analysis is based on Bray-Curtis distance calculated using the nematodes' normalised inferred diets. If an organism is consumed by a particular type of nematode then it is plotted in the vicinity of the region representing that nematode type.

The NMDS analysis was repeated with the Dikarya phyla removed from the dataset and Figures 5.12 and 5.13 show the results. The procedure returned a stress factor of 0.123 which again signifies that the figures are very good representations of the multidimensional data. With the potential nematophages removed, the main NMDS plot (Figure 5.12) appears different. The type 2B nematodes are more closely grouped with one outlier and there now seems to be more of a distinction between nematodes of type 1B and 2A with the 1B nematodes closer to the upper right portion of the plot and the 2A nematodes further down and to the left. The only type 1A nematode is, again, isolated from the rest.

Table 5.5 shows the results of eight permutation ANOVA tests for significance on normalised and rarefied datasets, some with/without Dikarya OTUs included and with either 2 or 4 feeding types defined. In all but one instance a significant p-value ($p < 0.05$) was returned while the rarefied dataset without Dikarya and with 2 feeding types defined gave a somewhat significant p-value of 0.059. The high residual R^2 values suggest that a lot of the variation is unexplained by the data.

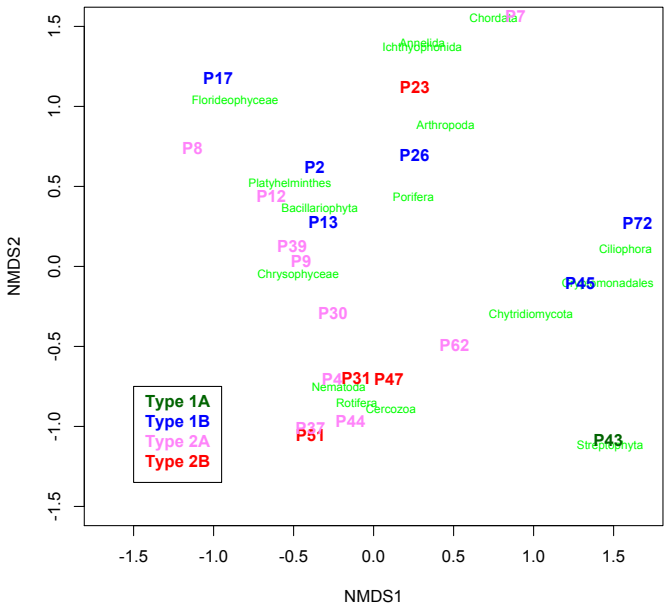

Figure 5.12: NMDS plot for individual nematodes based on Bray-Curtis distance calculated using their normalised inferred diets with Dikarya OTUs removed. Points representing nematodes are plotted close to points representing their diet.

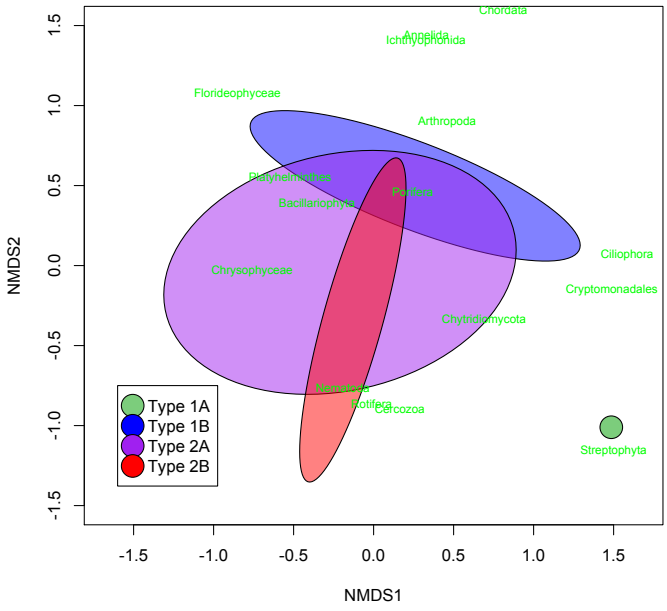

Figure 5.13: NMDS plot for individual nematodes showing the spread of nematodes of each Wieser feeding type. The analysis is based on Bray-Curtis distance calculated using the nematodes' normalised inferred diets with Dikarya OTUs removed. If an organism is consumed by a particular type of nematode then it is plotted in the vicinity of the region representing that nematode type.

Normalised/Rarefied	Dikarya Included?	No. Feeding Types	Residual R^2	p-value
Normalised	Yes	4	0.775	0.024*
Normalised	Yes	2	0.897	0.022*
Normalised	No	4	0.803	0.047*
Normalised	No	2	0.904	0.045*
Rarefied	Yes	4	0.789	0.045*
Rarefied	Yes	2	0.898	0.037*
Rarefied	No	4	0.799	0.059.
Rarefied	No	2	0.906	0.046*

Table 5.5: Results of permutation ANOVA to test whether feeding type can be inferred by diet. For the rarefied data, random subsamples of 20 were taken - the analysis was repeated 10 times and mean values were used. For the data with two feeding types, nematodes of Wieser feeding type 1A were combined with those of type 1B and type 2A nematodes were combined with type 2B nematodes. An asterisk (*) denotes a significant p-value ($p < 0.05$) and a full stop (.) denotes a somewhat significant p-value ($p < 0.1$).

5.8.1 Discussion

It is difficult to explain the apparent lack of a relationship between the assigned feeding types and the inferred interactions from both datasets. Table 5.4 summarises several statistics that could possibly be influenced by the feeding type of the species. Trophic levels, calculated from the inferred food web, should obviously be higher for predators and, additionally, the diversity of an organisms diet may be expected to be wider for deposit feeders than for predators. None of the statistics in Table 5.4 show any clear correlation with their assigned feeding types.

From Figures 5.6 to 5.13, it can be seen that the inferred diets are different for each feeding type and the results seem to follow what would be expected - for example, nematodes make up a large amount of the diet of predators (type 2B). The NMDS analyses suggest that nematodes of feeding types 1B and 2B are specialised feeders with a wider range of diet for those with feeding type 2A. It is harder to assess the results from type 1A nematodes because there was only one sample available to analyse.

The high residual R^2 values from the permutation ANOVA tests also indicate that much of the variation in the results are not explained by the feeding type data. This hidden variance is probably attributable to differences in the nematode's environment and community composition as, clearly, the nematode can only eat what food is available to it. Some nematodes of the same species may also vary in size which will have an impact on what they are able to consume. It is additionally important to consider that, just because part of a nematode's potential diet wasn't sequenced, it doesn't mean that the nematode doesn't eat it. The organism concerned may have not been picked up by sequencing, or may simply not have been present at the time of sequencing.

A narrower range of diet makes sense for predators (2B) and a wider range for type 2A

also agrees with its feeding type description. However, it is difficult to ascertain whether a nematode's diet would necessarily be identifiable by its phyla. Because of the variable nature in size of the majority of meiofauna species, it is conceivable that some nematodes will consume whatever organic matter is of a manageable size and also present in their habitat. Indeed, the results seem to suggest that this is the case with type 2A nematodes. A further problem is that even the most widely used feeding type definitions (118) (119) are necessarily vague reflecting the difficulty of classifying feeding types in such small, diverse and abundant organisms and the uncertainty over the nematodes' adaptability to different sources of food. In addition to this, it is unfortunate that it was only possible to classify 8 of the nematodes to species level using LCA classifier, otherwise this information could also have been used to assign feeding types to known species. However, the variability in size and behaviour of different nematodes species that share higher level taxonomic levels meant that the LCA classifications could not be used for this purpose.

Figures 5.6 to 5.13 along with the results from the permutation ANOVA tests provide evidence that feeding type can be inferred from data such as those from the experiment on individual nematodes but, because of the the variation of diet within feeding types, there will always be some uncertainty about which feeding types new data should be assigned. One quick and imprecise method of doing this would be to plot a new data point on the NMDS plot and examine which feeding type(s) it appears to best align with. A more mathematically thorough method is to perform a multinomial logistic regression using the *multinom* and *predict* functions in **R** to return the probabilities of new observations belonging to each feeding type. This approach was used on data for the 46 nematodes for which it was not possible to assign a feeding type and, of these, 5 were predicted to be type 1A, 12 were predicted to be 1B, 13 were predicted to be type 2A and 14 were predicted to be type 2B. The remaining 2 were not classified because they had no inferred diet belonging to the phyla that were present in the original 23 nematodes' collective inferred diets.

Despite the reported difficulties, the results are promising and there is evidence that nematodes of different feeding types have significantly different diets to each other. However, rather than concluding that it is possible to assign feeding types to nematodes based on their diet, a better interpretation of the results may be that although there is much variation in the diets of nematodes, even those of the same feeding type, it is possible to infer what a nematode has eaten by sequencing the nematode. This data may depend on other factors than a nematode's species, such as its size, its habitat and the community composition of this habitat. Thus, any future studies into predator-prey relationships within communities of small organisms may wish to consider this method of inferring these relationships.

In the future, it would be interesting to combine the 18S sequencing of nematodes with 16S sequencing because many of the nematodes with tiny mouths, such as those of type 1A, are known, or assumed, to be microvores. Bacteria will make up a large part of the diets of these nematodes and this data is missing from a study that only uses 18S sequencing. In this study only one individual nematode of each species was sequenced, further studies may benefit from the sequencing of multiple individuals of the same species - from the same/different location(s) - to investigate the variation in the diets of worms of the same species.

5.9 Results - Comparing the Community and Individual Datasets

To test the validity of the food web shown in Figure 5.4 it is required to compare it with other matrices generated from the community dataset. Various matrices were constructed, as described in Sections 5.3 and 5.4, to form interaction networks. After some OTUs were removed due to the pre-processing criteria outlined in Section 5.4.1 and all OTUs which did not belong to both datasets were filtered out, 17 OTUs remained. The various matrices of interaction networks were reduced in size accordingly and were ready for comparison. No similarities between these matrices were immediately apparent.

Four different ROC analyses were carried out, each using a different interaction network as the gold standard. Two used feeding interaction networks generated from the experiments on individual nematodes, the f and I interaction matrices, and two used interaction networks generated from the co-occurrence data, the L1 Precision and SparCC matrices. The reason for this was to investigate different scenarios because it is not possible to be sure of the true interactions using the available data.

Figures 5.14, 5.15, 5.16 and 5.17 show some of the results obtained using the direct effort matrix (f), the indirect effort matrix (I), the L1 precision matrix and the SparCC matrix, respectively as the gold standard. The areas under these ROC curves (AUROC) are shown in Figures 5.18, 5.19, 5.20 and 5.21. When the f and I matrices were used as the gold standard, the AUROC values for the L1 Precision, Pearson correlation and Spearman correlation matrices were the highest at around 0.65. When the L1 precision matrix was used, the Bray-Curtis dissimilarity and Kulback-Leibler divergence matrices returned higher AUROC values but the values for the f and I matrices were both around 0.5. For the final choice of gold standard, the SparCC matrix, only the Bray-Curtis dissimilarity matrix produced an AUROC value of higher than 0.6. The f and I matrices both produced values of less than 0.5 in this instance.

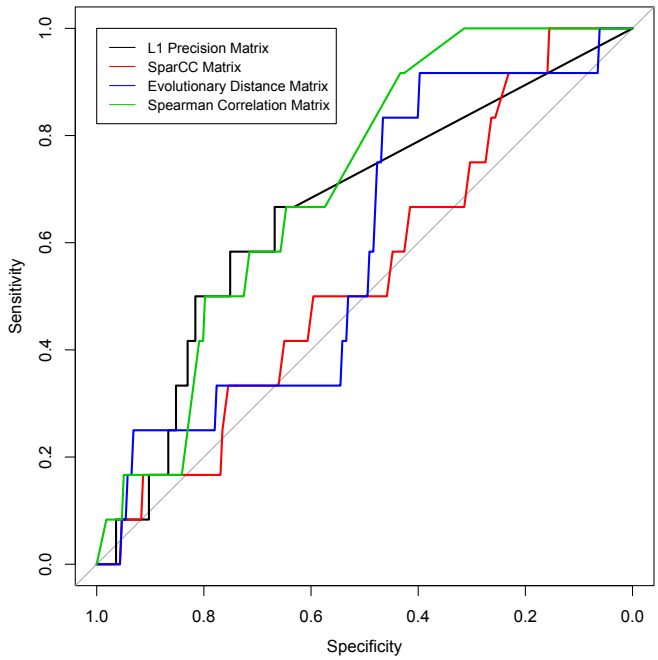

Figure 5.14: ROC analysis to assess the effectiveness of various similarity measures, using f matrix as gold standard. The f matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

Method	Direct Efforts	Indirect Efforts	SparCC	L1 Prec.	Pearson	Spearman	Bray-Curtis	Kulback-Leibler
Indirect Efforts	1.000	-	-	-	-	-	-	-
SparCC	0.974	0.974	-	-	-	-	-	-
L1 Precision	0.958	0.913	0.949	-	-	-	-	-
Pearson Correlation	0.968	1.000	0.909	0.935	-	-	-	-
Spearman Correlation	0.941	0.941	0.870	0.912	0.429	-	-	-
Bray-Curtis Dissimilarity	0.929	0.966	0.905	0.893	0.417	0.385	-	-
Kulback-Leibler Divergence	1.000	1.000	0.956	1.000	1.000	1.000	1.000	-
Evolutionary Distance	0.933	1.000	1.000	1.000	0.947	0.952	0.944	0.946

Table 5.6: Jaccard distances between interaction network graphs generated from different matrices.

Pairwise Jaccard distances between all nine interaction networks are shown in Table 5.6. The numbers are predominantly high and reveal that few of the edges, representing interactions, were shared between graphs and that the majority of inferred interaction networks differed widely from each other. The exceptions to this were the graphs generated from the Bray-Curtis, Pearson and Spearman matrices which shared more than half of their interactions.

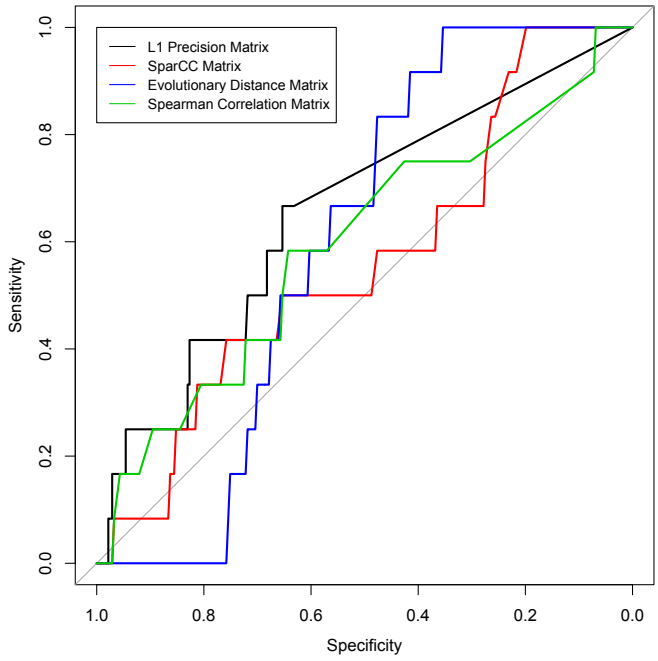

Figure 5.15: ROC analysis to assess the effectiveness of various similarity measures, using I matrix as gold standard. The I matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

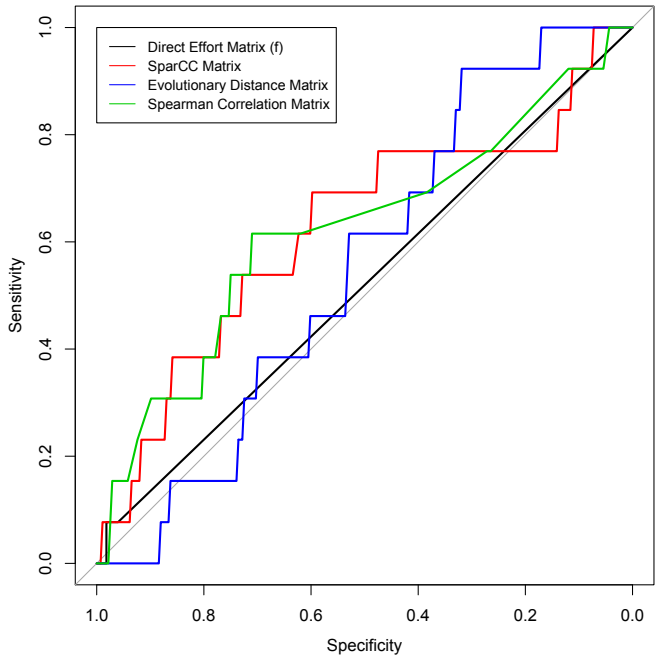

Figure 5.16: ROC analysis to assess the effectiveness of various similarity measures, using L1 precision matrix as gold standard. The L1 precision matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

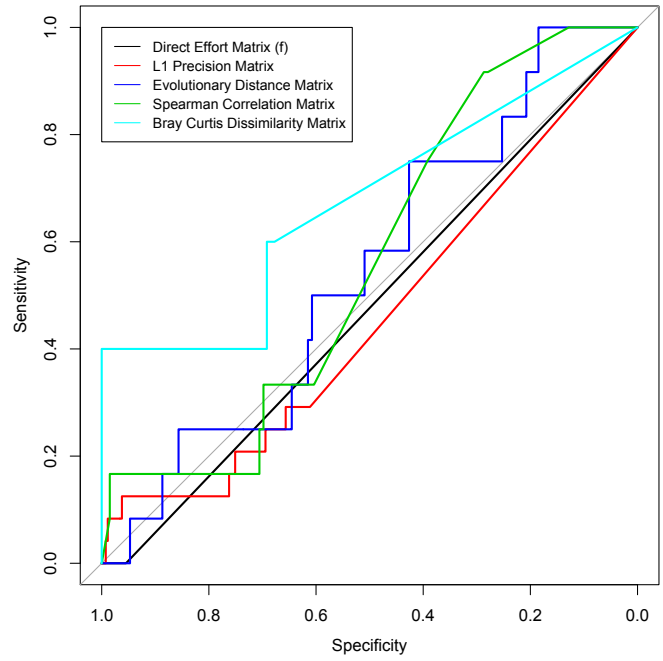

Figure 5.17: ROC analysis to assess the effectiveness of various similarity measures, using SparCC matrix as gold standard. The SparCC matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

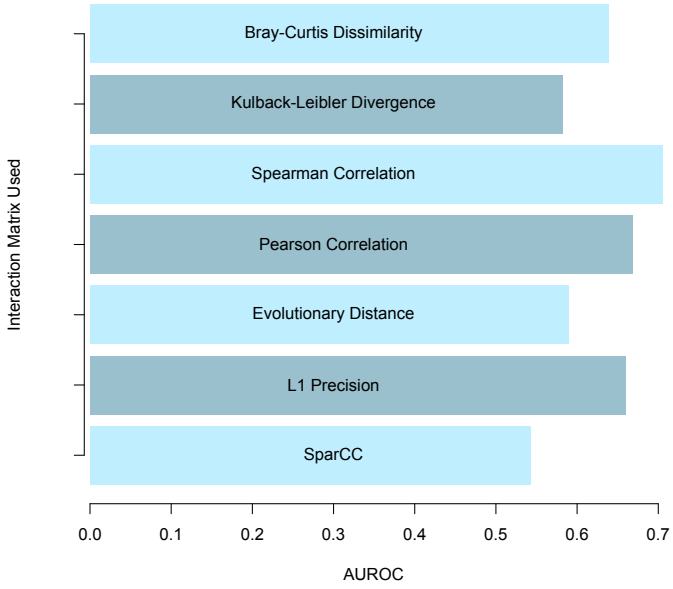

Figure 5.18: AUROC for various similarity measures when using the direct effort matrix (f) as gold standard. The f matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

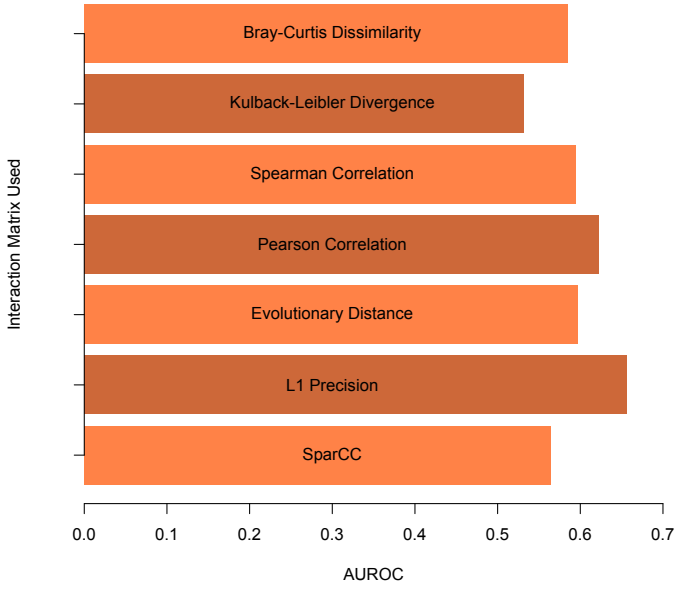

Figure 5.19: AUROC for various similarity measures when using the indirect effort matrix (I) as gold standard. The I matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

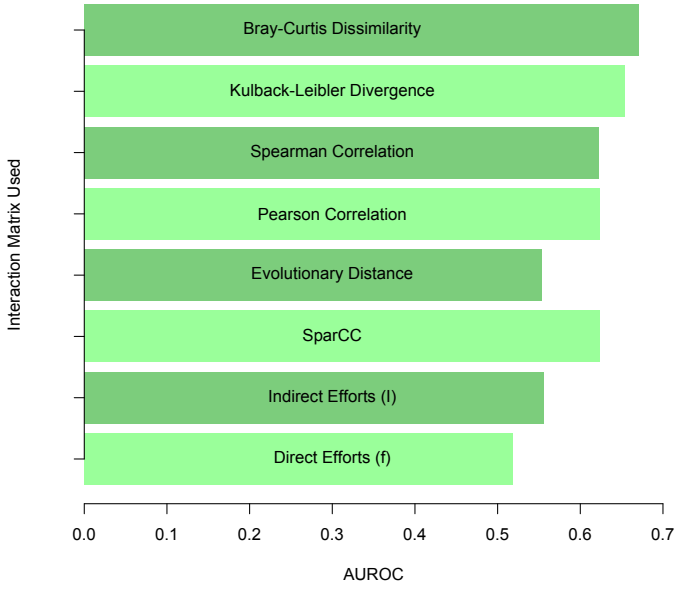

Figure 5.20: AUROC for various similarity measures when using the L1 precision matrix as gold standard. The L1 precision matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

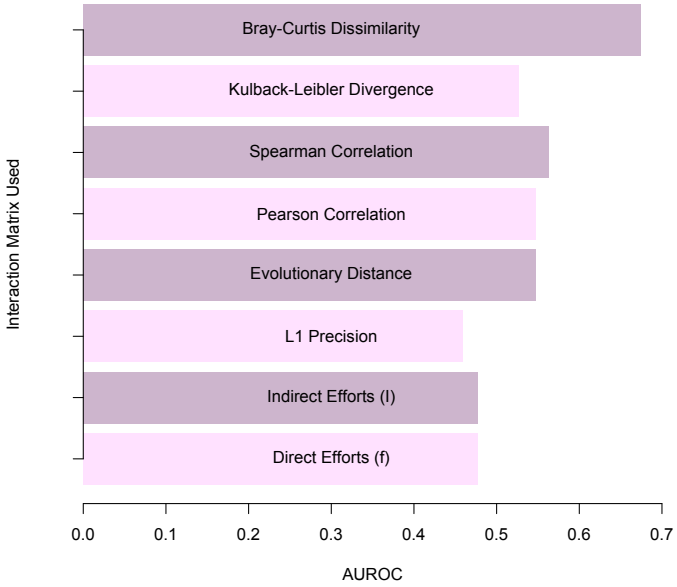

Figure 5.21: AUROC for various similarity measures when using the SparCC matrix as gold standard. The SparCC matrix is taken to be the true matrix of interactions - true positives (sensitivity) and false positives (specificity) from the other data are found by comparing with this gold standard.

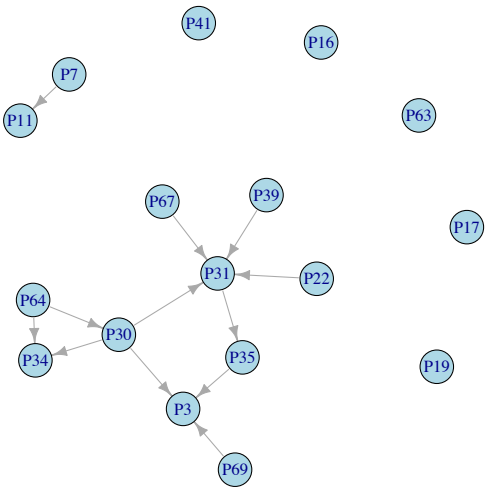

Figure 5.22: Food web graph generated from the matrix of direct efforts (f) between the 17 OTUs occurring in both the f graph and interaction network graphs generated from co-occurrence data.

Edge	P7 → P11	P22 → P31	P30 → P3	P30 → P31	P30 → P34	P31 → P35
Degree of Prey OTU	1	1	4	4	4	5
Degree of Predator OTU	1	5	3	5	2	2
Betweenness Centrality of Prey OTU	0.0	0.0	18.0	18.0	18.0	22.5
Betweenness Centrality of Predator OTU	0.0	22.5	9.5	22.5	0.0	4.0
Closeness Centrality of Prey OTU	3.906×10^{-3}	7.092×10^{-3}	7.519×10^{-3}	7.519×10^{-3}	7.519×10^{-3}	7.519×10^{-3}
Closeness Centrality of Predator OTU	3.906×10^{-3}	7.519×10^{-3}	7.299×10^{-3}	7.519×10^{-3}	7.143×10^{-3}	7.299×10^{-3}
Presence in Co-occurrence Graphs	3/6	0/6	3/6	0/6	0/6	0/6
Edge	P35 → P3	P39 → P31	P64 → P30	P64 → P34	P67 → P31	P69 → P3
Degree of Prey OTU	2	1	2	2	1	1
Degree of Predator OTU	3	5	4	2	5	3
Betweenness Centrality of Prey OTU	4.0	0.0	0.0	0.0	0.0	0.0
Betweenness Centrality of Predator OTU	9.5	22.5	18.0	0.0	22.5	9.5
Closeness Centrality of Prey OTU	7.299×10^{-3}	7.092×10^{-3}	7.143×10^{-3}	7.143×10^{-3}	7.092×10^{-3}	6.897×10^{-3}
Closeness Centrality of Predator OTU	7.299×10^{-3}	7.519×10^{-3}	7.519×10^{-3}	7.143×10^{-3}	7.519×10^{-3}	7.299×10^{-3}
Presence in Co-occurrence Graphs	0/6	0/6	1/6	1/6	0/6	0/6

Table 5.7: Analysis of edges representing feeding interactions in the food web graph that was generated from the matrix of direct efforts (f) between OTUs occurring in both the f graph and interaction network graphs generated from co-occurrence data. The final row shows the number of different co-occurrence generated graphs, out of 6, in which a particular edge is present.

Figure 5.22 shows a subset of the causal food web graph (the food web generated from the f matrix) which contains only the OTUs which were also present in the other interaction network graphs investigated in this section (those generated using co-occurrence data). Because many of the edges present in this graph were not present in the interaction networks that were generated from co-occurrence data, it was decided to examine certain properties of the nodes which these edges joined together. The properties chosen were the degree, betweenness centrality and closeness centrality of each node. The results of this investigation are presented in Table 5.7 and it can be seen that two of the edges appeared in half of the co-occurrence graphs but the properties of these edges were not similar to each other. There was no obvious distinction between these edges and those that were not present in any of the co-occurrence graphs.

The values for the degree at each of the shared 17 OTU nodes in each of the nine interaction network graphs are displayed in Figure 5.23 and the number of corresponding nodes sharing the same degree for each pair of graphs is shown in Table 5.8 (e.g. OTU P3 has a degree of five in both the indirect interaction graph and the SparCC graph). These results again show a reasonably high similarity between the Pearson correlation, Spearman correlation and Bray-Curtis dissimilarity graphs with nine or ten out of the seventeen nodes sharing the same degree. All other graphs have fewer than half of their nodes sharing the same degree with another graph's corresponding nodes, although there is some similarity between the Spearman correlation graph and the evolutionary distance graph and also between the indirect effort matrix and the L1 Precision matrix for which eight corresponding nodes share the same degree in each case.

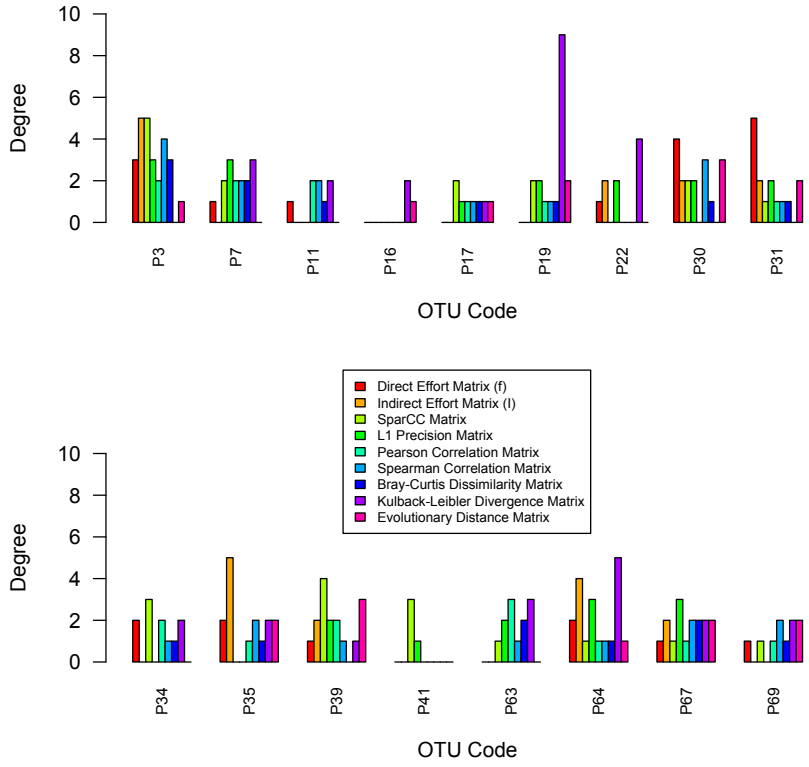

Figure 5.23: Degree at each OTU node on interaction network graphs generated from different matrices.

Method	Direct Efforts	Indirect Efforts	SparCC	L1 Prec.	Pearson	Spearman	Bray-Curtis	Kulback-Leibler
Indirect Efforts	5	-	-	-	-	-	-	-
SparCC	3	4	-	-	-	-	-	-
L1 Precision	2	8	5	-	-	-	-	-
Pearson Correlation	5	3	7	3	-	-	-	-
Spearman Correlation	4	3	6	2	9	-	-	-
Bray-Curtis Dissimilarity	5	3	6	4	10	10	-	-
Kulback-Leibler Divergence	4	2	0	2	6	7	3	-
Evolutionary Distance	3	7	4	5	4	8	5	5

Table 5.8: Number of corresponding OTU nodes with the same degree on interaction network graphs generated form different matrices.

	Direct Efforts	Indirect Efforts	SparCC	L1 Prec.	Pearson	Spearman	Bray-Curtis	Kulback-Leibler	Evol. Distance
Betweenness Centrality	3.176	0.941	5.118	8.294	0.118	1.882	1.706	4.471	0.059
Closeness Centrality	5.782×10^{-3}	4.816×10^{-3}	6.399×10^{-3}	7.010×10^{-3}	4.025×10^{-3}	4.829×10^{-3}	4.479×10^{-3}	9.216×10^{-3}	4.000×10^{-3}

Table 5.9: Mean values for the betweenness centrality and closeness centrality of nine interaction network graphs generated using different interaction matrices.

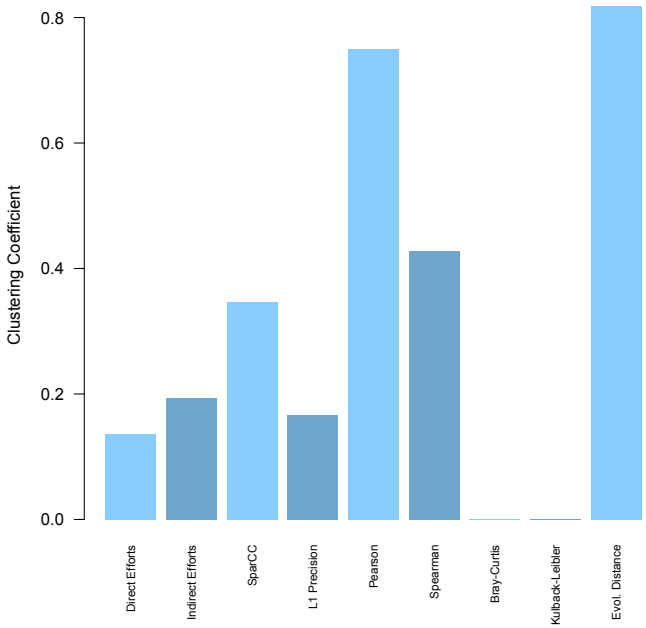

Figure 5.24: Clustering coefficients for interaction network graphs generated from different matrices.

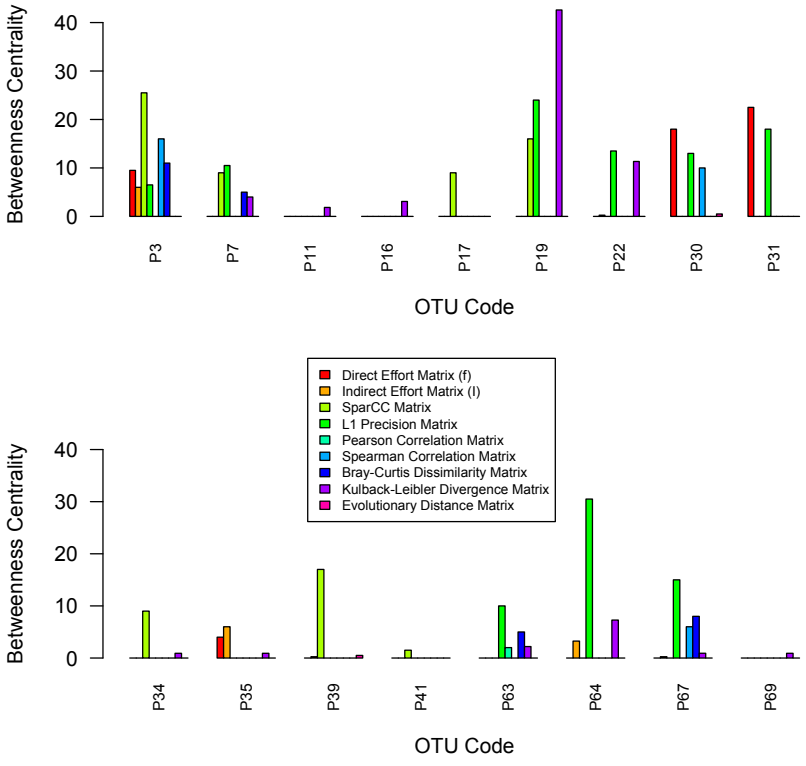

Figure 5.25: Betweenness centrality at each OTU node on interaction network graphs generated from different matrices.

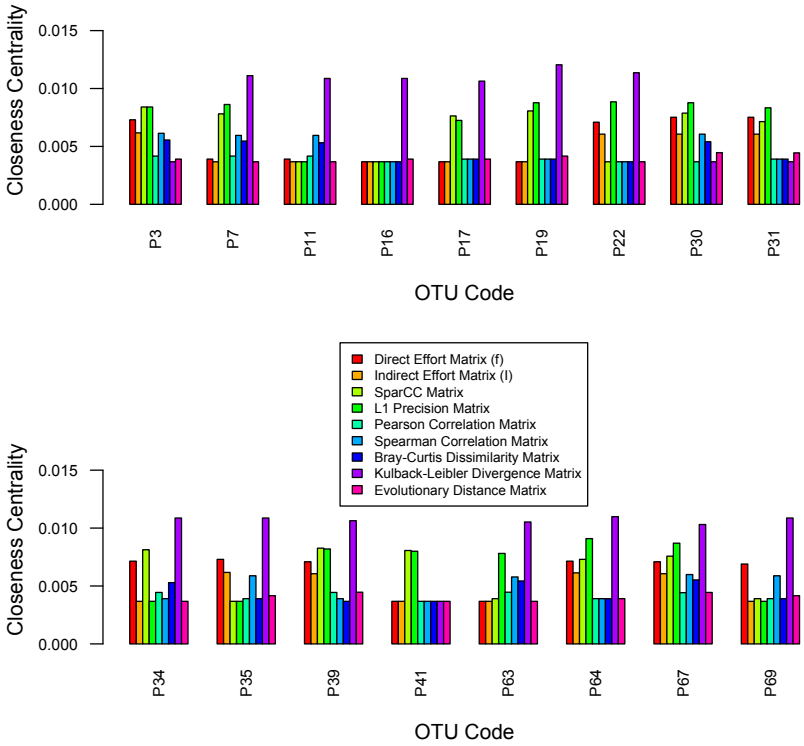

Figure 5.26: Closeness centrality at each OTU node on interaction network graphs generated from different matrices.

Figure 5.24 shows the clustering coefficient for each of the interaction network graphs. The evolutionary distance graph and the Pearson correlation graph showed the highest level of clustering whereas the Bray-Curtis dissimilarity graph and the Kulback Leibler divergence graph both had a clustering coefficient of zero, indicating that these graphs contained no closed triplets.

The betweenness centrality for nodes in all interaction network graphs are shown in Figure 5.25. It can be seen that some OTUs are important, in this respect, in a number of different graphs, most notably OTU P3 which has a relatively high betweenness centrality in six out of the nine graphs (all apart from the Pearson correlation graph, the Kulback-Leibler divergence graph and the Evolutionary distance graph). Some nodes, such as P11, P16, P41 and P69 have low values in all of the graphs, suggesting that these OTUs are often inactive. There is much variation in the mean values of betweenness centrality, as presented in Table 5.9, with the L1 Precision graph showing a high value, suggesting that many of its nodes are often active. Conversely, the Evolutionary distance graph has a low mean betweenness centrality which suggests a low level of interaction between most of its nodes.

In contrast to the betweenness centrality values, the values for closeness centrality (Figure 5.26) appear fairly uniform within each graph and across each node. The figure, in conjunction with the mean values shown in Table 5.9, shows that the OTU nodes in the Kulback-Leibler divergence graph have consistently higher closeness centrality values than the nodes in other graphs. Corresponding nodes in the SparCC graph and the L1 Precision graph tend to have similar values, and those in the remaining graphs tend to have lower values that are similarly distributed.

5.9.1 Discussion

The ROC analysis indicates a poor level of prediction regardless of which interaction network was used as the gold standard, although a large majority of the AUROC values are greater than 0.5 which does suggest that the values in the various interaction matrices are slightly better than completely random data in each case.

When f and I were used as the gold standard, the ROC curves for the Pearson correlation, Spearman correlation and L1 Precision data returned higher values, suggesting that these measures are better for detecting feeding interactions. The AUROC results were marginally higher when f was used as the gold standard, suggesting that that these methods are slightly better at detecting direct predator-prey interactions than indirect ones. The evolutionary tree based approach shows a slightly better consensus with both f and I matrices. There are

more likely to be feeding relationships between distantly related species than closely related species. This agrees with the expectation that closely related species are more likely to be competing for the same resources.

When interaction networks generated from co-occurrence data (L1 Precision and SparCC) were used as the gold standard, AUROC values generated from the f and I matrices indicated that, if the gold standards were accurate, the f and I matrices did not predict the presence of interactions.

These results present a dilemma. If the causal food web generated from the f matrix is a good representation of true interactions then methods which use co-occurrence data to infer interactions are ineffective, at least on this kind of data. The reverse also holds; if the co-occurrence generated interaction networks are a realistic gold standard then analysis using the f and I matrices is ineffective. However, the findings in Section 5.8 have provided evidence that the individual dataset (Section 5.2.1) can be used to determine a nematode's diet and this, consequently, provides evidence that a food web generated from these data is legitimate. Following on from this, it can be concluded that the f food web is a reasonable choice for gold standard and that the co-occurrence data is not useful for inferring feeding interactions for data of this type.

The lack of matching edges between the causal food web graph and other interaction networks shown in Table 5.7 further demonstrates that the co-occurrence approach failed to define the same interactions as those present in the f matrix. There also seems to be no pattern involving the properties of the nodes that surround the edges that do appear in co-occurrence generated graphs and those that don't. Again, this shows that the causal food web differs greatly from the interaction networks generated from co-occurrence data and supplies evidence that these methods are not detecting feeding interactions.

Overall, there is little evidence of any agreement between the relationships inferred from the two datasets considered in this report and there are a number of reasons why this could be. It could be the case that some or all of the extra OTUs found in the single nematode experiments were not in fact ingested by the nematodes in question and have ended up in the samples by some other means. Foreign organic material could, perhaps, have been transferred onto the exterior of the worms as a result of contact between two organisms. However, as stated, this is unlikely because of the results in Section 5.8.

Another reason for this lack of consensus may be due to the difficulty of defining the nature of predator-prey relationships when using a matrix of correlations derived from co-occurrence

data. For a relationship between two species there will be either a positive or a negative value for their correlation, however, a predator-prey interaction is more complex. There are actually two directed interactions representing the negative effect of the presence of predator on the abundance of the prey and the positive effect of the presence of the prey on the abundance of the predator. Note that correlation matrices are symmetric and therefore can only show a single, undirected, interaction between two species. As argued, this is insufficient to describe predator-prey relationships.

A third explanation is that, maybe, both methods are correctly detecting predator-prey interactions but the results derived from the co-occurrence data are overwhelmed by the detection of many other different types of interaction in such a way that the presence of the predator-prey interactions is obscured.

Among the findings in Chapter 2 is evidence that meiofauna diversity is mainly driven by niche overlap. For such communities, the use of co-occurrence data may not be the best approach for inferring interactions, particularly feeding relationships, because the abundances of species that share a niche will be strongly correlated regardless of the presence of interactions. The results in Chapter 2 also show that, for the more dominant meiofauna phyla (Nematodes and Platyhelminthes), community composition is similar in samples with similar environmental and geographical characteristics, specifically sediment grain size, seawater surface temperature, distance between samples and latitude. The fact that this impacts the distribution of these phyla will also limit the effectiveness of co-occurrence data when used to determine interactions. It is probable that for studies where co-occurrence is less dependent on so many factors, the use of such data will work better.

In summary, the analysis of the individual nematode data yielded good evidence that feeding relationships can be inferred using these data which casts doubt upon the validity of other methods. The lack of corroboration between the results from the two datasets suggests that it is possible that the methods currently being used on co-occurrence data are inappropriate in some cases, particularly when applied to the inference of feeding relationships within communities where the composition is dependent on a range of environmental factors. This is clearly of importance because it could lead to invalid conclusions in studies that choose to apply these methods. However, it should be emphasised that these conclusions can only be applied to data of the type investigated in this chapter and that co-occurrence data has been shown to be capable of detecting interactions between species in more favourable datasets (120).

Chapter 6

Discussion

The work presented in this thesis shows that DNA sequencing data obtained from a small number of experiments can be used for a wide range of analyses, some of them directly related to investigating the structure and nature of the source of the sample (Chapters 2 and 5) and others focusing on improving the methods used for processing the resultant data (Chapters 3 and 4). See Figure 6.1.

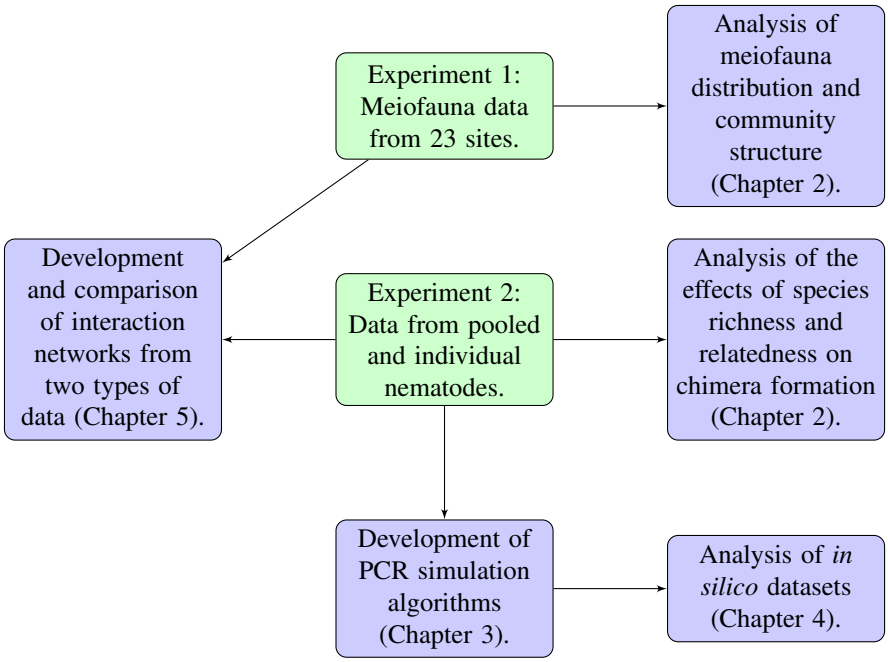

Figure 6.1: Analysis carried out using sequencing data from two experiments.

6.1 Summary of Analysis and Results

Chapter 2 is split into two parts, each of which describes a different experiment. The first experiment (45) is a survey of meiofaunal communities at sites around Europe and one in

Africa. This reveals greater diversity than previously expected, increased correlation in community between geographically closer sites, and suggests that meiofauna are controlled by niche effects whereas protozoa are more consistent with the neutral model. It was also shown, using rarefaction, that sampling effort was generally insufficient and sequencing methods which can generate higher number of reads than 454 pyrosequencing, such as Illumina sequencing by synthesis, would be desirable for similar studies. Although more numerous reads would improve the understanding of sample richness, some uncertainty would still remain. It is inevitable that, regardless of the number of reads, some OTUs will still be missed. Additionally, the presence of noisy reads and the possibility of sample contamination will artificially inflate the estimated richness of a sample (126) (127), reinforcing the need for good noise removal software and good experimental practice.

The second part of Chapter 2 describes an experiment on pooled nematodes (46) which were divided into samples of phylogenetically close and distant species. The results from these experiments show that richer, more distantly related samples tended to produce a greater number of chimeras and that the chimera break points were more likely to occur at areas of lower nucleotide diversity.

It is noted in Chapter 3 that existing PCR simulation software is inadequate, particularly in relation to the simulated chimeras. This provided motivation for the development of the Simera and Simera 2 PCR simulation algorithms which were shown to produce realistic chimeric sequences. The Simera 2 algorithm was used as part of the analysis of *in silico* datasets in Chapter 4 which highlighted problems with chimera removal software that were, until now, largely unnoticed. Of particular concern was the observation that substantially fewer chimeras were being detected than had been previously believed. Conclusions drawn from the results of *in silico* microbial community analysis were somewhat worrying due to the levels of uncertainty that were revealed in these results and that were caused, in part, by the presence of unwanted chimeras.

Chapter 5 concludes that a nematode's diet, and hence its feeding type, can be inferred from sequencing the DNA of the individual nematode in question, suggesting that this method is an appropriate way of determining feeding interactions. Other methods of predicting interactions between nematodes, using co-occurrence data, were shown to yield networks with little or no similarity to those found using the individual nematode data. From this it could be concluded that these methods are unreliable when used to infer feeding relationships from similar data.

6.2 DNA Sequencing and the Future

NGS platforms already allow a seemingly endless amount of research possibilities. A few of these, which have been covered in this thesis, include microbial community analysis, development of simulation software and the inference of interaction networks. Other areas include medical applications, the analysis of drinking water and sewage treatment facilities and the development of personal cleaning products.

As DNA sequencing technology becomes cheaper and increases in performance and accessibility, and with the impending introduction of new third generation sequencing platforms, research potential can only increase in scope. However, it is important that the development of associated bioinformatic tools does not lag too far behind. As rapid advancements in technology occur, a wide variety of different data must be analysed and, because of this, the software used to process and analyse sequencing data will quickly become suboptimal and eventually obsolete. New software, designed to work in conjunction with new technology and its associated data must constantly be developed if meaningful conclusions are desired from future research.

Further computational challenges will also arise with the inevitable increase in the size of datasets possible from sequencing with new technologies. For example, the Illumina HiSeq 2000 machine, released in 2010, can reportedly yield 1000 times as many reads as the 454 FLX Titanium with 1 billion reads possible for the former versus 1 million reads for the latter which was released two years earlier. Comparing the HiSeq 2000 with its predecessor, the GAIIx, shows that the potential number of reads has increased more than threefold, from 320 million to 1 billion (5).

Good data processing and noise removal strategies are clearly important, then, but these can only be effective if they are partnered with experimental methods which minimise the amount of errors. An increase in errors is sometimes acceptable because of a huge increase in throughput, as was the case with the introduction of NGS technologies, but it is important not to compound these errors by using poor experimental protocols.

Differing methods of PCR amplification have been shown to affect the quality of results. For example, it has been shown that an increased number of PCR rounds will increase the quantity of chimeric sequences in the output (46) (76) (78) (74). It is important to avoid sample contamination so that unwanted DNA is not amplified. Products generated from previous PCR amplifications are a common source of contamination and one way to combat this is the segmentation of laboratories in such a way as to avoid this happening. PCR preparation and

post-reaction processing can be carried out in separate areas (128). Determining the optimal concentration of Magnesium to be used as a co-factor for the DNA polymerase is also of importance (129). If the concentration is too low then the activity of the polymerase will be reduced (130). If the concentration is too high then this can lead to the stabilisation of double stranded DNA which can inhibit denaturation of DNA during the reaction (129) (130).

Roux (131) outlines a strategy which suggests a trial and error approach to PCR optimisation. Initially starting with primer pairs with similar melting temperatures, using a range of Mg^{++} concentrations and $10^4 - 10^5$ copies of the template, various alterations in the PCR conditions are suggested based on the quality of the output.

An interesting development contributing to the increasing levels of accessibility and affordability of DNA sequencing is the imminent introduction of the portable *MinION* sequencer (132) from Nanopore Technologies. The MinION will cost around £600 (133) and can be plugged directly into a computer via USB, making DNA sequencing available to a large number of individual users for the first time. Reported error rates are high (133), meaning that the practical uses of this technology are currently limited. Nevertheless, if improvements can be made, the implications for the future are very interesting. On the spot sequencing will be possible and this will, for example, be very useful in the field of medicine because samples will be available for sequencing immediately, thus saving time and reducing the risk of contamination. The technology will also prove useful for the analysis of samples from remote areas and other locations where access to sequencing technology would otherwise be difficult, such as sites in developing countries.

To summarise, whilst the advancements and achievements made in recent decades are impressive, research presented in this thesis and elsewhere has shown that NGS output is often error-strewn and, therefore, the degree of confidence in the interpretation of results must be reduced. It is of high importance that the bioinformatic tools available keep the accuracy of sequencing output at at least the current level and it would be preferable that further improvements are made. Any amount of sequencing data is of no use if it cannot be communicated reliably.

6.3 Chimeras and Noise Removal

Since the beginning of the use of next generation sequencing there has been awareness of chimeras as a problematic source of noise (126) (18). The findings in this thesis have suggested that, nevertheless, their importance has been underestimated, mainly due to the inability of chimera detection software to effectively eliminate chimeras from data. These results

were found from analysis of realistically diverse *in silico* datasets and are contrary to what was believed from testing software on less realistic mock community data (18) (20).

As mentioned in Chapter 4, less effective chimera removal from sequencing data will naturally lead to a more pessimistic appraisal of the reliability of any analysis based on the use of technology from which chimera formation is possible. In order to restore confidence in the conclusions found from next generation sequencing analysis of microbial communities it will be necessary to improve chimera removal software. Fortunately, similar *in silico* datasets to those which were used to discover the deficiencies may be of use in testing and perfecting future chimera detection software. The development of the PCR simulation algorithms presented in Chapter 3 was an integral step towards the subsequent generation of the aforementioned *in silico* datasets.

The investigation into the drivers of chimera formation in Chapter 2 provided evidence that chimera distributions are not random and can be predicted based on the collection of sequences representing the DNA from which they are formed (46). This information, and the chimera abundances found from the same experiment, were used to develop the Simera 1 and Simera 2 algorithms which were shown to create more realistic chimeras than other PCR simulators. There is incentive to further improve the performance of these algorithms by using a range of different datasets to select the most appropriate parameters. This would further improve the quality of simulated *in silico* datasets which would, in turn, allow better testing of chimera detection software.

Future analysis into chimeras, their causes, their simulation and their detection is heavily reliant on which technologies will be used for future DNA sequencing. Third generation sequencing methods eliminate the need for PCR amplification so chimera formation will not be an issue if they are used. However, the widespread use of third generation sequencing platforms is still a number of years away and chimera formation is still a very serious problem associated with Illumina, the most common NGS platform. Therefore, improved chimera detection in future studies is paramount. If this is not achievable then an awareness of the limitations of the results obtained from NGS is necessary.

Whilst the research in this thesis has concentrated mainly on chimeras as a source of noise, the analysis of *in silico* datasets in Chapter 4 showed that sequencing noise and PCR errors also adversely affected the reliability of results when analysing community structure. There are two possible ways of reducing noise from sequencing data in the future. The first, and most desirable, is the development of low-noise or even noise-free sequencing platforms and the second is to develop more effective noise removal software. No assumptions can be

made about the former, so continued research on noise removal is necessary. Adequate noise removal has been demonstrated for 454 pyrosequencing data using AmpliconNoise (18) and Illumina data using QIIME (21) but, to reiterate, it is important that new software is developed in order to match any new technology that is used.

6.4 Community Analysis

The use of NGS data for the analysis of meiofaunal and microbial communities has been incredibly useful in a huge number of studies, including the study of meiofauna at different marine benthic locations around Europe and Africa presented in the first part of Chapter 2. A large number of interesting results were gained from this experiment which demonstrate the usefulness of DNA sequencing and metagenomics in the analysis of small organisms.

In addition to this, the analysis of *in silico* communities presented in Chapter 4 has shown that the information yielded from similar analysis can be compared against the known community structure in order to determine how accurate the derived metrics are. The findings from Chapter 4 suggest that the richness and diversity estimates for the communities investigated may be higher than the true values.

6.5 Interaction Prediction

The techniques employed in Chapter 5 to generate a food web from sequencing data showed good results when compared to methods using co-occurrence data, and should provoke interesting research using similar methods. As has already been noted in Section 5.9.1 this does not mean that more widely used methods can, or should, be discredited.

The data that were used to generate interaction networks in this thesis came from meiofauna communities in marine benthic regions, mainly around the coast of Europe. As discussed in Chapter 2, the distribution and diversity of the organisms in these communities is driven by niche overlap and environmental factors. The discussion in Section 5.9.1 explains that if these effects are strong enough then they can override any correlation in co-occurrence caused by species interactions and, consequently, it is not possible to use co-occurrence based methods to detect feeding relationships in such datasets.

Co-occurrence based methods have been shown to be effective in previous studies where the data have been more favourable (36) and the limitations have been noted (120). In the future, it would be interesting to compare these different methods of interaction inference

on communities where co-occurrence based methods have been shown to be successful. For now, the method of diet inference from sequencing data offers an alternative approach, especially in cases for which other approaches struggle.

6.6 Publications and Future Work

The analysis of the two experiments presented in Chapter 2 has been published (46) (45) and there are also plans to publish the PCR simulation and *in silico* dataset analysis and to make the simulation software available.

Additional sequencing of nematode DNA from the 18S gene has taken place ahead of further analysis of feeding interactions, this has been carried out in conjunction with sequencing of the 16S gene from the same samples in order to identify the bacteria associated with each nematode sample. If this analysis produces positive results then it will provide great encouragement to perform similar studies to infer feeding relationships using sequencing data.

There is also much potential to use the PCR simulation software to generate *in silico* datasets for the testing of chimera detection software. Indeed, these methods have already been used for the testing of the "UCHIME" option in the development of *VSEARCH* (134), an open source tool that aims to emulate and, in some areas, improve USEARCH which is not open source.

Appendix A

Appendix to Chapter 3: Probability Distributions

This appendix describes the probability distributions that were used to generate the random variables involved in the models developed in Chapter 3.

A.1 The Binomial Distribution

The binomial distribution is a discrete probability distribution that shows the probable number of successes out of n independent true or false trials with each trial having a fixed probability, p, of success. These independent yes/no experiments are known as $Bernoulli\ trials$.

The probability of x successes is

$$\Pr(X = x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where $n \in \{1, 2, \dots\}$; $x \in \{0, 1, \dots, n\}$; $0 \leq p \leq 1$. The binomial distribution has mean np and variance $np(1-p)$.

A.2 The Multinomial Distribution

The multinomial distribution is the multivariate generalisation of the binomial distribution. Instead of n trials with two outcomes (true/false Bernoulli trials) there are now k possible outcomes for each of the n trials, each with its own probability of success, $p_1, p_2, \dots, p_k$ such that $p_1 + p_2 + \cdots + p_k = 1$.

The distribution has the probability mass function

$$\Pr(X_1 = x_1; \ldots; X_k = x_k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

where $n, k \in \{1, 2, \ldots\}$; $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$; $x_1 + x_2 + \cdots + x_k = n$.

A.3 The Multivariate Hypergeometric Distribution

The multinomial distribution can be thought of as the probability distribution which describes drawing n differently coloured balls, with k different colours, from an urn with replacement. The multivariate hypergeometric distribution, in contrast, is the probability distribution which describes drawing n differently coloured balls from the same urn $without$ replacement.

For a hypergeometric distribution with k coloured balls there are N balls in total with N_1 balls of colour 1, ... and N_k balls of colour k such that $N_1 + N_2 + \cdots + N_k = N$. The parameter $n \leq N$ again represents the number of balls to be drawn from the urn. The multivariate hypergeometric distribution has the probability mass function

$$\Pr(X_1 = x_1; \ldots; X_k = x_k) = \frac{\prod_{i=1}^{k} \binom{N_i}{x_i}}{\binom{N}{n}}$$

where $N, k \in \{1, 2, \ldots\}$; $n \in \{1, 2, \ldots, N\}$; $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$; $x_1 + x_2 + \cdots + x_k = n$.

Note that the univariate hypergeometric distribution is simply a special case of the multivariate hypergeometric distribution where $k = 2$.

A.4 Wallenius' Multivariate Non-central Hypergeometric Distribution

Continuing the analogy of drawing coloured balls from an urn, Wallenius' multivariate non-central hypergeometric distribution can be thought of as the probability distribution which describes drawing n differently coloured balls, with k different colours, from an urn without replacement, as with the multivariate hypergeometric distribution. However, in this case, each different colour has a different weight associated with it. Balls with higher weights

are more likely to be selected, meaning that the colour of the ball that is drawn depends not only on its quantity but also on its weighting. Typically these weights are allowed to take any positive real number value but, for convenience, may be normalised so that their sum is equal to 1.

Therefore, in addition to the parameters included in the multivariate hypergeometric distribution, this distribution requires another vector of parameters, $\omega_1, \ldots, \omega_k$, giving the k different weights of each differently coloured ball. The probability mass function of Wallenius' multivariate non-central hypergeometric distribution is given by

$$\Pr(X_1 = x_1; \ldots; X_k = x_k) = \left(\prod_{i=1}^{k} \binom{N_i}{x_i} \right) \int_0^1 \prod_{i=1}^{k} \left(1 - t^{\omega_i / D}\right)^{x_i} \mathrm{d}t$$

where

$$D = \sum_{i=1}^{k} \omega_i (N_i - x_i)$$

and $N, k \in \{1, 2, \ldots\}$; $n \in \{1, 2, \ldots, N\}$; $x_1, \ldots, x_k \in \{0, 1, \ldots, n\}$; $x_1 + x_2 + \cdots + x_k = n$; $\omega_1, \ldots, \omega_k > 0$.

Bibliography

[1] F. Sanger, S. Nicklen, and A. R. Coulsen, "DNA Sequencing with Chain-Terminating Inhibitors," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 5463–5467, 1977.

[2] "Polymerase chain reaction," Encyclopedia Britannica: http://www.britannica.com/EBchecked/topic/468736/polymerase-chain-reaction, 2014, accessed: 24/2/2015.

[3] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, ..., and J. M. Rothberg, "Genome Sequencing in Microfabricated High-Density Picolitre Reactors," *Nature*, vol. 437, pp. 326–7, 2005.

[4] J. G. Caporaso, C. L. Lauber, W. A. Walters, D. Berg-Lyons, C. A. Lozupone, P. J. Turnbaugh, N. Fierer, and R. Knight, "Global Patterns of 16S rRNA Diversity at a Depth of Millions of Sequences per Sample," *Proceedings of the National Academy of Sciences of the United States of America*, 2010.

[5] "Illumina website," http://www.illumina.com/, accessed: 21/2/2015.

[6] "Life Technologies website," http://www.ilifetechnologies.com/, accessed: 21/2/2015.

[7] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, ..., and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5910, pp. 133–138, 2009.

[8] J. H. Lee, E. R. Daugharthy, J. Scheiman, R. Kalhor, T. C. Ferrante, R. Terry, ..., and G. M. Church, "Fluorescent in situ sequencing (FISSEQ) of RNA for gene expression profiling in intact cells and tissues," *Nature Protocols*, vol. 10, pp. 442–458, 2015.

[9] E. R. Mardis, "Next-generation DNA sequencing methods," *Annual Reviews*, vol. 9, pp. 387–402, 2008.

[10] J. F. Thompson and K. E. Steinmann, "Single molecule sequencing with a HeliScope genetic analysis system," *Current Protocols in Molecular Biology, Chapter 7:Unit7.10. doi: 10.1002/0471142727.mb0710s92*, 2010.

[11] A. L. Laszlo, I. M. Derrington, B. C. Ross, H. Brinkerhoff, A. Adey, I. C. Nova, ..., and J. H. Gundlach, "Decoding long nanopore sequencing reads of natural DNA," *Nature Biotechnology*, vol. 32, pp. 829–833, 2014.

[12] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, ..., and S. Turner, "Real-time DNA sequencing from single polymerase molecules," *Science*, vol. 323, no. 5920, pp. 133–138, 2009.

[13] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[14] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, pp. 195–197, 1981.

[15] H. T. T. M. Kazutaka Katoh, Kei-ichi Kuma, "MAFFT version 5: improvement in accuracy of multiple sequence alignment," p. 511518, 2005. [Online]. Available: http://nar.oxfordjournals.org/content/33/2/511

[16] M. A. Larkin, G. Blackshields, ..., and D. G. Higgins, "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23 (21), pp. 2947–2948, 2007.

[17] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[18] C. Quince, A. Landzen, and P. Turnbaugh, "Removing noise from pyrosequenced amplicons," *BMC Bioinformatics*, vol. 12, 2011.

[19] Homer, *Iliad*, 760–710 BC, 6.179–182.

[20] R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, and R. Knight, "UCHIME improves sensitivity and speed of chimera detection," *Bioinformatics doi: 10.1093/bioinformatics/btr381*, 2011.

[21] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, ..., and R. Knight, "QIIME allows analysis of high-throughput community sequencing data," *Nature Methods*, vol. 7, pp. 335–336, 2010.

[22] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, ..., and C. F. Weber, "Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities," *Applied and Environmental Microbiology*, vol. 75, no. 23, pp. 7537–7541, 2009.

[23] S. Creer, V. G. Fonseca, D. L. Porazinska, R. M. Giblin-Davis, W. Sung, D. M. Power, ..., and W. K. Thomas, "Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promises," *Molecular Ecology*, vol. 19 (Suppl. 1), pp. 4–20, 2010.

[24] D. L. Porazinska, R. M. Giblin-Davis, L. Faller, W. Farmerie, N. Kanzaki, K. Morris, ..., and W. K. Thomas, "Evaluating high-throughput sequencing as a method for metagenomic analysis of nematode diversity," *Molecular Ecology Resources*, vol. 9, pp. 1439–1450, 2009.

[25] T. H. Jukes and C. R. Cantor, *Evolution of Protein Molecules*. New York: Academic Press, 1969.

[26] A. Chao, "Nonparametric estimation of the number of classes in a population," *Scandinavian Journal of Statistics, Theory and Applications*, vol. 11, no. 4, pp. 265–270, 1984.

[27] K. P. Burnham and W. S. Overton, "Robust estimation of population size when capture probabilities vary among animals," *Ecology*, vol. 60, no. 5, pp. 927–936, 1979.

[28] E. P. Smith and G. van Belle, "Nonparametric estimation of species richness," *Biometrics*, vol. 40, pp. 119–129, 1984.

[29] J. Hortal, P. A. V. Borges, and C. Gaspar, "Evaluating the performance of species richness estimators: sensitivity to sample grain size," *Journal of Animal Ecology*, vol. 75, pp. 274–287, 2006.

[30] A. Chao and S. Lee, "Estimating the number of classes via sample coverage," *Journal of the American Statistical Association*, vol. 87, no. 417, pp. 210–217, 1992.

[31] C. E. Shannon and W. Weaver, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423 and 623–656, 1948.

[32] R. H. Whittaker, "Vegetation of the Siskiyou Mountains, Oregon and California," *Ecological Monographs*, vol. 30, pp. 279–338, 1960.

[33] E. C. Pielou, "The measurement of diversity in different types of biological collections," *Journal of Theoretical Biology*, vol. 13, pp. 131–144, 1966.

[34] J. R. Bray and J. T. Curtis, "An ordination of the upland forest communities of southern Wisconsin," *Ecological Monographs*, vol. 27, no. 4, pp. 325–349, 1957.

[35] G. W. Snedecor, *Calculation and Interpretation of Analysis of Variance and Covariance*. Collegiate Press, Ames, Iowa, 1934.

[36] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower, "Co-occurrence relationships in the human microbiome," *PLoS Comput Biol 8(7): e1002606. doi:10.1371/journal.pcbi.1002606*, 2012.

[37] K. Pearson, "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. 58, pp. 240–242, 1895.

[38] C. E. Spearman, "The proof and measurement of association between two things," *American Journal of Psychology*, vol. 15, pp. 72–101, 1904.

[39] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79–86, 1951.

[40] J. Freedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS Computational Biology*, vol. 8(9): e1002687. doi:10.1371/ journal.pcbi.1002687, 2012.

[41] Q. Ruan, D. Dutta, M. S. Schwalbach, J. A. Steele, J. A. Fuhrman, and F. Sun, "Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors," *Bioinformatics*, vol. 22, pp. 2532–2538, 2006.

[42] K. C. Li, "Genome-wide coexpression dynamics: Theory and application," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 16 875–16 880, 2002.

[43] S. P. Hubbell, *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, 2001.

[44] J. H. Vandermeer, "Niche theory," *Annual Review of Ecology and Systematics*, vol. 3, pp. 107–132, 1972.

[45] V. G. Fonseca, G. R. Carvalho, B. Nichols, C. Quince, H. F. Johnson, S. P. Neill, ..., and S. Creer, "Metagenetic analysis of patterns of distribution and diversity of marine meiobenthic eukaryotes," *Global Ecology and Biogeography*, vol. 23, no. 11, pp. 1293–1302, 2014.

[46] V. G. Fonseca, B. Nichols, D. Lallias, C. Quince, G. R. Carvalho, D. M. Power, and S. Creer, "Sample richness and genetic diversity as drivers of chimera formation in nssu metagenetic analyses," *Nucleic Acids Research; doi: 10.1093/nar/gks002*, 2012.

[47] O. Giere, *Meiobenthology: the microscopic motile fauna of aquatic sediments*, 2nd ed. Springer, Berlin, 2009.

[48] L. Zinger, L. A. Amaral-Zettler, J. A. Fuhrman, M. C. H.-D. ans S. M. Huse, D. B. M. Welch, ..., and A. Ramette, "Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems," *PLoS ONE, 6, e24570*, 2011.

[49] V. G. Fonseca, G. R. Carvalho, W. Sung, H. F. Johnson, D. M. Power, S. P. Neill, ..., and S. Creer, "Second-generation environ- mental sequencing unmasks marine metazoan biodiversity," *Nature Communications, 1, doi: 10.1038*, 2010.

[50] T. Stoeck, D. Bass, M. Nebel, R. Christen, M. D. M. Jones, H. W. Breiner, and T. A. Richards, "Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water," *Molecular Ecology*, vol. 19, pp. 21–31, 2010.

[51] H. M. Bik, W. Sung, P. De Ley, J. G. Baldwin, J. Sharma, A. Rocha-Olivares, and W. K. Thomas, "Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments," *Molecular Ecology*, vol. 21, no. 5, pp. 1048–1059, 2012.

[52] R. Logares, S. Audic, D. Bass, L. Bittner, R. Christen, J. M. Claverie, ..., and R. Massana, "Patterns of rare and abundant marine microbial eukaryotes," *Current Biology*, vol. 24, pp. 813–821, 2014.

[53] H. M. Platt and R. M. Warwick, *Free-living marine nematodes. Part II: British chromadorids*. Synopses of the British Fauna (new series) 38. Brill, Leiden, 1988.

[54] "DEFRA website," http://chartingprogress.defra.gov.uk, accessed: 21/2/2015.

[55] "NOAA website," http://www.ncdc.noaa.gov/, accessed: 21/2/2015.

[56] M. Yoder, I. T. De Ley, I. W. King, M. Mundo-Ocampo, J. Mann, M. Blaxter, L. Poiras, and P. De Ley, "DESS: a versatile solution for preserving morphology and extractable DNA of nematodes," *Nematology*, vol. 8, pp. 367–376, 2006.

[57] R. K. Colwell, "EstimateS: statistical estimation of species richness and shared species from samples," University of Connecticut, Storrs, CT. Available at http://viceroy.eeb.uconn.edu/estimates/, 2013.

[58] K. Clarke and R. Gorley, "PRIMER v6: user manual/tutorial," PRIMER-E, Plymouth, 2006.

[59] W. R. Rice, "Analyzing tables of statistical tests," *Evolution*, vol. 43, pp. 223–225, 1988.

[60] K. Harris, T. Parsons, U. Z. Ijaz, L. Lahti, I. Holmes, and C. Quince, "Linking statistical and ecological theory: Hubbell's unified neutral theory of biodiversity as a hierarchical Dirichlet process," *Proceedings of the IEEE, Accepted for Publication*, 2015.

[61] B. J. Finlay, "Global dispersal of free-living microbial eukaryote species," *Science*, vol. 296, pp. 1061–1063, 2002.

[62] J. A. Gilbert, J. A. Steele, J. G. Caporaso, L. Steinbrück, J. Reeder, B. Temperton, ..., and D. Field, "Defining seasonal marine microbial community dynamics," *ISME Journal*, vol. 6, no. 2, pp. 298–308, 2012.

[63] B. Chen, L. Zheng, B. Huang, S. Song, and H. Liu, "Seasonal and spatial comparisons of phytoplankton growth and mortality rates due to microzooplankton grazing in the northern South China Sea," *Biogeosciences*, vol. 10, pp. 2775–2785, 2013.

[64] R. H. Findlay and L. Watling, "Seasonal variation in the structure of a marine benthic microbial community," *Microbial Ecology*, vol. 36, pp. 23–30, 1998.

[65] J. L. Green, A. J. Holmes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Billings, and A. Beattie, "Spatial scaling of microbial eukaryote diversity," *Nature*, vol. 432, pp. 747–750, 2004.

[66] R. C. Pitcher, P. Lawton, N. Ellis, S. J. Smith, L. S. Incze, C. L. Wei, ..., and P. V. R. Snelgrove, "Exploring the role of environmental variables in shaping patterns of seabed biodiversity composition in regional-scale ecosystems," *Journal of Applied Ecology*, vol. 49, pp. 670–679, 2012.

[67] W. Foissner, "Biogeography and dispersal of microorganisms: a review emphasizing protists," *Acta Protozoologica*, vol. 45, pp. 111–136, 2006.

[68] C. Mora, D. P. Tittensor, S. Adl, A. G. B. Simpson, and B. Worm, "How many species are there on Earth and in the ocean?" *PLoS Biology*, vol. 9, e1001127, 2011.

[69] M. J. Costello, P. Bouchet, C. S. Emblow, and A. Legakis, "European marine biodiversity inventory and taxonomic resources: state of the art and gaps in knowledge," *Marine Ecology Progress Series*, vol. 316, pp. 257–268, 2006.

[70] P. J. D. Lambshead and G. Boucher, "Marine nematode deep-sea biodiversity – hyperdiverse or hype?" *Journal of Biogeography*, vol. 30, pp. 475–485, 2003.

[71] J. A. Huber, D. B. M. Welch, H. G. Morrison, S. M. Huse, P. R. Neal, D. A. Butterfield, and M. L. Sogin, "Microbial population structures in the deep marine biosphere," *Science*, vol. 318, pp. 97–100, 2007.

[72] M. L. Sogin, H. G. Morrison, J. A. Huber, D. B. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl, "Microbial diversity in the deep sea and the underexplored "rare biosphere"," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, pp. 12 115–12 120, 2006.

[73] R. Massana and C. Pedros-Alio, "Unveiling new microbial eukaryotes in the surface ocean." *Current Opinion in Microbiology*, vol. 11, pp. 213–218, 2008.

[74] B. J. Haas, D. Gevers, A. M. Earl, M. Feldgarden, D. V. Ward, G. Giannoukos, ..., and B. W. Birren, "Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons," *Genome Research*, vol. 21, pp. 494–504, 2011.

[75] K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman, "New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras." *Applied and Environmental Microbiology*, vol. 72, pp. 5734–5741, 2006.

[76] G. C. Wang and Y. Wang, "Frequency of formation of chimeric molecules as a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes," *Applied and Environmental Microbiology*, vol. 63, pp. 4645–4650, 1997.

[77] F. von Wintzingerode, U. B. Gobel, and E. Stackebrandt, "Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis," *FEMS Microbiology Reviews*, vol. 21, pp. 213–229, 1997.

[78] G. C. Wang and Y. Wang, "The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species," *Microbiology*, vol. 142, pp. 1107–1114, 1996.

[79] B. H. Meldal, N. J. Debenham, P. De Ley, I. T. De Ley, J. R. Vanfleteren, A. R. Vierstraete, ..., and P. J. D. Lambshead, "An improved molecular phylogeny of the Nematoda with special emphasis on marine taxa," *Molecular Phylogenetics and Evolution*, vol. 42, pp. 622–636, 2007.

[80] K. Tamura, J. Dudley, M. Nei, and S. Kumar, "Molecular evolutionary genetics analysis (mega) software version 4.0," *Molecular Biology and Evolution*, vol. 24, pp. 1596–1599, 2007.

[81] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," *Methods in Molecular Biology*, vol. 453, pp. 3–31, 2008.

[82] X. Qiu, L. Wu, H. Huang, P. E. McDonel, A. V. Palumbo, J. M. Tiedje, and J. Zhou, "Evaluation of PCR-generated chimeras, mutations, and heteroduplexes with 16S rRNA gene-based cloning," *Applied and Environmental Microbiology*, vol. 67, pp. 880–887, 2001.

[83] R. P. Smyth, T. E. Schlub, A. Grimm, V. Venturi, A. Chopra, S. Mallal, M. P. Davenport, and J. Mak, "Reducing chimera formation during PCR amplification to ensure accurate genotyping," *Gene*, vol. 469, pp. 45–51, 2010.

[84] J. Reeder and R. Knight, "The 'rare biosphere': a reality check," *Nature Methods*, vol. 6, pp. 636–637, 2009.

[85] S. Creer, "Second-generation sequencing derived insights into the temporal biodiversity dynamics of freshwater protists," *Molecular Ecology*, vol. 19, pp. 2829–2831, 2010.

[86] R. A. Clayton, G. Sutton, P. S. Hinkle Jr., C. Bult, and C. Fields, "Intraspecific variation in small-subunit rRNA sequences in GenBank: why single sequences may not adequately represent prokaryotic taxa," *International Journal of Systematic Bacteriology*, vol. 45, pp. 595–599, 1995.

[87] E. Rubin and A. A. Levy, "A mathematical model and a computerized simulation of PCR using complex templates," *Nucleic Acids Research*, vol. 24, pp. 3538–3545, 1996.

[88] G. F. Ficetola, E. Coissac, S. Zundel, T. Riaz, W. Shehzad, J. Bessière, P. Tarberiet, and F. Pompanon, "An in silico approach for the evaluation of DNA barcodes," *BMC Genomics*, vol. 11, p. 434, 2010.

[89] S. Wu and U. Manber, "Agrep - a fast approximate pattern-matching tool," in *In Proc. of USENIX Technical Conference*, 1992, pp. 153–162.

[90] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, pp. 707–710, 1965.

[91] "cybertory.org PCR simulator," http://cybertory.org/simulators/pcr/index.html, accessed: 21/2/2015.

[92] "bioinformatics.org PCR simulator," http://www.bioinformatics.org/sms2/pcr_products.html, accessed: 21/2/2015.

[93] "amnh.org PCR simulator," http://www.amnh.org/learn/pd/genetics/pcr/pcr_directions.html, accessed: 21/2/2015.

[94] W. A. Walters, J. G. Caporaso, C. L. Lauber, D. Berg-Lyons, N. Fierer, and R. Knight, "PrimerProspector: de novo design and taxonomic analysis of PCR primers," *Bioinformatics*, vol. 27 (8), pp. 1159–1161, 2011.

[95] F. E. Angly, D. Willner, F. Rohwer, P. Hugenholtz, and G. W. Tyson, "Grinder: a versatile amplicon and shotgun sequence simulator," *Nucleic Acids Research*, vol. 40 (12), 2012.

[96] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[97] R. C. Edgar, "Search and clustering orders of magnitude faster than BLAST," *Bioinformatics*, vol. 26 (19), pp. 2460–2461, 2010.

[98] R. G. Rutledge and D. Stewart, "A kinetic-based sigmoidal model for the polymerase chain reaction and its application to high-capacity absolute quantitative real-time PCR," *BMC Biotechnology 8:47 doi:10.1186/1472-6750-8-47*, 2008.

[99] A. Fog, "Pseudo random number generators," http://www.agner.org/random/, accessed: 21/2/2015.

[100] T. A. Arnold and J. W. Emerson, "Nonparametric goodness-of-fit tests for discrete null distributions," *The R Journal*, vol. 3, no. 2, pp. 34–39, 2011.

[101] S. Balzer, K. Malde, A. Lanzén, and I. Jonassen, "Characteristics of 454 pyrosequencing data – enabling realistic simulation with flowsim," *Bioinformatics*, vol. 26 (18), pp. 1420–1425, 2010.

[102] B. J. Haas, "Chimera Slayer reference database," http://microbiomeutil.sourceforge.net/, accessed: 21/2/2015.

[103] Q. Wang, "RDP classifier training database," http://sourceforge.net/projects/rdp-classifier/files/RDP_Classifier_TrainingData/, accessed: 21/2/2015.

[104] "Full Greengenes 16S unaligned fasta database – last updated 9/5/2011," http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/, accessed: 21/2/2015.

[105] "Full Silva 16S unaligned fasta database – last updated 9/5/2011," http://greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/, accessed: 21/2/2015.

[106] F. W. Preston, "The commonness, and rarity, of species," *Ecology*, vol. 29 (3), pp. 254–283, 1948.

[107] J. Dunbar, S. M. Barns, L. O. Ticknor, and C. R. Kuske, "Empirical and theoretical bacterial diversity in four arizona soils," *Applied and Environmental Microbiology*, vol. 68, no. 6, pp. 3035–3045, 2002.

[108] S. S. Hirano, E. V. Nordheim, D. C. Arny, and C. D. Upper, "Lognormal distribution of epiphytic bacterial populations on leaf surfaces," *Applied and Environmental Microbiology*, vol. 44, no. 3, pp. 695–700, 1982.

[109] J. E. Loper, T. V. Suslow, and M. N. Schroth, "Log-normal distribution of bacterial populations in the rhizosphere," *Phytopathology*, vol. 74, pp. 1454–1460, 1984.

[110] T. P. Curtis, W. T. Sloan, and J. W. Scannel, "Estimating prokaryotic diversity and its limits," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 16, pp. 10 494–10 499, 2002.

[111] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, ..., and J. I. Gordon, "A core gut microbiome in obese and lean twins," *Nature*, vol. 457, pp. 480–484, 2009.

[112] P. J. D. Lambshead, "Recent developments in marine benthic biodiversity research," *Oceanis*, vol. 19 (6), pp. 5–24, 1993.

[113] C. L. Nunn, V. O. Ezenwa, C. Arnold, and W. D. Koenig, "Mutualism or parasitism? Phylogentic approach to characterize the Oxpecker-Ungulate relationship," *Evolution*, vol. 65, pp. 1297–1304, 2011.

[114] J. M. Willey, L. M. Sherwood, and C. J. Woolverton, *Prescott's Microbiology*, 9th ed. McGraw-Hill Higher Education, 2013.

[115] C. Quince, P. G. Higgs, and A. J. McKane, "Deleting species from model food webs," *OIKOS*, vol. 110 (2), pp. 283–296, 2005.

[116] E. Pruesse, C. Quast, K. Knittel, B. Fuchs, W. Ludwig, J. Peplies, and F. O. Glockner, "SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB," *Nucleic Acids Research*, vol. 35, pp. 7188–7196, 2007.

[117] A. Lanzén, S. L. Jørgensen, D. Huson, M. Gorfer, S. H. Grindhaug, I. Jonassen, L. Øvreås, and T. Urich, "CREST classification resources for environmental sequence tags," *PLos ONE*, vol. 7(11): e49334. doi:10.1371/journal.pone.0049334, 2012.

[118] W. Wieser, "Die Beziehung zwischen Mundhöhlengestalt, Ernährungsweise und Vorkommen bei freilebenden marinen Nematoden," *Arkiv für Zoologie*, vol. 2, pp. 439–484, 1953.

[119] T. Moens and M. Vincx, "Observations on the feeding ecology of estuarine nematodes," *Journal of the Marine Biological Association of the United Kingdom*, vol. 77 (01), pp. 211–227, 1997.

[120] D. Berry and S. Widder, "Deciphering microbial interactions and detecting keystone species with co-occurrence networks," *Frontiers in Microbiology*, vol. 5:219, 2014.

[121] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont, "Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data," *Journal of Machine Learning Research*, vol. 9, pp. 485–516, 2008.

[122] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix," *Molecular Biology and Evolution 26:1641-1650, doi:10.1093/molbev/msp077*, 2009.

[123] P. Jaccard, "Étude comparative de la distribution florale dans une portion des Alpes et du Jura," *Bulletin de la Société Vaudoise des Sciences Naturelles*, vol. 37, pp. 547–579, 1901.

[124] G. Csardi and T. Nepusz, "The igraph software package for complex network research," *InterJournal*, vol. Complex Systems, p. 1695, 2006. [Online]. Available: http://igraph.org

[125] H. Anke, M. Stadler, A. Mayer, and O. Sterner, "Secondary metabolites with nematicidal and antimicrobial activity from nematophagous fungi and Ascomycetes," *Canadian Journal of Botany*, vol. 73 (suppl. 1): S932–S939, 1995.

[126] C. Quince, A. Landzen, T. P. Curtis, R. J. Davenport, N. Hall, I. M. Head, L. F. Read, and W. T. Sloan, "Accurate determination of microbial diversity from 454 pyrosequencing data," *Nature Methods*, vol. 6, pp. 639–41, 2009.

[127] M. J. Morgan, A. A. Charlton, D. M. Hartley, L. N. Court, and C. M. Hardy, "Improved inference of taxonomic richness from environmental DNA," *PLoS ONE*, vol. 8(8): e71974. doi:10.1371/journal.pone.0071974, 2013.

[128] B. J. Balin, H. C. Gérard, E. J. Arking, D. M. Appelt, P. J. Branigan, J. T. Abrams, J. A. Whittum-Hudson, and A. P. Hudson, "Identification and localization of chlamydia pneumoniae in the Alzheimer's brain," *Medical Microbiology and Immunology*, vol. 187, pp. 23–42, 1998.

[129] P. Markoulatos, N. Siafakas, and M. Moncany, "Multiplex polymerase chain reaction: a practical approach," *Journal of Clinical Laboratory Analysis*, vol. 16, pp. 47–51, 2002.

[130] "Promega: nucleic acid amplification protocols and guidelines," http://www.promega.co.uk/resources/product-guides-and-selectors/protocols-and-applications-guide/pcr-amplification/, accessed: 25/3/2015.

[131] K. M. Roux, "Optimization and troubleshooting in PCR," *Cold Spring Harbor Protocols,* doi:10.1101/pdb.ip66, 2009.

[132] "The Nanopore MinION," https://nanoporetech.com/technology/the-minion-device-a-miniaturised-sensing-system/the-minion-device-a-miniaturised-sensing-system, accessed: 25/3/2015.

[133] E. C. Hayden, "Nanopore genome sequencer makes its debut," *Nature,* doi:10.1038/nature.2012.10051, 2012.

[134] T. Rognes, F. Mahé, T. Flouri, B. Nichols, U. Z. Ijaz, and C. Quince, "VSEARCH: a versatile open source metagenomics tool," *To be published*, 2015.

[135] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2010, ISBN 3-900051-07-0. [Online]. Available: http://www.R-project.org/